Approximating Bayes: Inference and Modeling

by

Jun Yang

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Statistical Sciences
University of Toronto

Abstract

Approximating Bayes: Inference and Modeling

Jun Yang
Doctor of Philosophy
Graduate Department of Statistical Sciences
University of Toronto
2020

This thesis focuses on sources of error in modern Bayesian analysis and machine learning in the "big data" era. We develop new theoretical tools for analyzing and controlling different sources of error. Our work leads to new theory and methodology for providing performance guarantees for modern Bayesian methods and machine learning algorithms.

The first two contributions of this thesis are new tools for studying the complexity/hardness of achieving approximation guarantees for Markov chain Monte Carlo (MCMC) in high-dimensional settings. The third contribution of this thesis is a theoretical framework for Bayesian analysis in the face of model misspecification that makes the analysis of different practical Bayesian methods possible. The fourth contribution of this thesis is a bridge between PAC-Bayes theory and Rademacher complexity, two important theoretical tools for developing the so-called generalization bounds in machine learning theory.

Through the analysis of various sources of error, this thesis makes contributions to the study of: convergence of high-dimensional MCMC; optimization of high-dimensional MCMC; Bayesian model misspecification; and statistical learning theory.

Contents

Chapter 1

Introduction

This thesis focuses on sources of error in modern Bayesian analysis and machine learning in the "big data" era. We develop new theoretical tools for analyzing and controlling different sources of error. Our work leads to new theory and methodology for providing performance guarantees for modern Bayesian methods and machine learning algorithms.

In modern Bayesian analysis, errors arise when we approximate Bayes rule in complex models using approximate inference algorithms such as Markov chain Monte Carlo (MCMC). Approximation is essential, but most algorithms cannot provide any strong guarantees on the accuracy of their answers without overly restrictive assumptions. The first two contributions of this thesis are new tools for studying the complexity/hardness of achieving approximation guarantees for MCMC in high-dimensional settings.

In Bayesian analysis, another source of error is *model misspecification*. Bayesian analysis implicitly assumes the model precisely captures all relevant aspects of the data. In practice, approximations in the model that lead to model misspecification can have a serious impact on Bayesian inference. The third contribution of this thesis is a theoretical framework for Bayesian analysis in the face of model misspecification that makes the analysis of different practical Bayesian methods possible.

In machine learning, overfitting is a source of error. In recent years, our understanding of overfitting has changed. Many modern machine learning algorithms are designed to perfectly fit the training data and so the challenge is to understand why overfitting in those models is "benign". We are interested in building new theoretical tools to study benign overfitting in modern machine learning algorithms. One challenge is that the performance of modern machine learning algorithms, such as deep neural networks, seems to hinge on favorable properties of the data and/or the training process. In this thesis, we study the *generalization* of machine learning algorithms, which is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data. The fourth contribution of this thesis is a bridge between PAC-Bayes theory and Rademacher complexity, two important theoretical tools for developing the so-called generalization bounds.

Through the analysis of various sources of error, this thesis makes contributions to the study of: convergence of high-dimensional MCMC; optimization of high-dimensional MCMC; Bayesian model misspecification; and statistical learning theory. In the remainder of this chapter, we provide a brief introduction to each research area, describe our contribution, and convey its significance to statistics and machine learning.

1.1 Analyzing the convergence of high-dimensional MCMC

Bayesian analysis relies on computations involving the so-called posterior distribution. However, for complex models, the analytical form of posterior distribution is usually not tractable. The development of Markov chain Monte Carlo (MCMC) methods made it possible to calculate numerical approximations of posterior distribution for complex models. Therefore, MCMC is now a key part of the foundation of modern Bayesian analysis and is widely used in practice.

The key idea of MCMC is to numerically simulate a Markov chain whose stationary distribution was designed to be the desired posterior distribution. The popularity of MCMC comes from the fact that, as the number of iterations grows, the distribution of the sample from MCMC asymptotically converges to the posterior distribution. In practice, there are two stages of running a MCMC algorithm. The first stage is called burn-in. In this stage, one runs MCMC for a finite number of steps and hopes the samples of the Markov chain afterwards can be viewed as approximate samples from the posterior distribution. After the burn-in stage, in the second stage, one collects multiple samples to compute approximations of quantities of interests.

Approximation errors in the burn-in stage arise since the samples after a finite number of iterations are only approximate. This type of errors is related to convergence properties of MCMC algorithms. The basic question is how many iterations the MCMC algorithms must be run in order to approximate posterior distribution to a desired precision. With the advent of big data in recent years, there is growing interest in the high-dimensional setting where the number of parameters of the statistical model is of the same (or even higher) order as the number of observations. The modern high-dimensional regime presents new challenges to our understanding of the convergence properties of MCMC algorithms. There is a lack of theoretical tools for convergence analysis of high-dimensional MCMC. Among practitioners, there is a general understanding that scaling classical MCMC algorithms to high dimension can be problematic and initialization can play a role in convergence.

In high-dimensional settings, one can understand the relationship between convergence and number of dimensions by studying "convergence complexity", which defines how the convergence time of MCMC algorithms scales with the dimension of the parameter set. In this thesis, we develop a new theoretical tool for obtaining tight complexity bounds for MCMC in high-dimensional settings. As a demonstration, we analyze a realistic Gibbs sampler and obtain a complexity bound for its convergence, which shows the MCMC algorithm scales well to high dimension, under certain conditions on the observed data and the initial state. This work presents evidence against the widely held belief that scaling MCMC to modern high-dimensional settings is not feasible. We expect the proposed tool can be applied to analyze convergence complexity of many other high-dimensional MCMC algorithms.

1.2 Optimizing high-dimensional Metropolis–Hastings algorithms

Approximation errors after the burn-in stage are Monte Carlo errors, which arise by using a finite number of samples to approximate the quantity of interests. Monte Carlo error exists even if the Markov chain has already converged to its stationary distribution. This type of errors is related to the efficiency of MCMC algorithms. To understand how to reduce this type of approximation errors, the basic question is how efficient the Markov chain can explore the state space provided that it has converged to the stationary distribution.

The optimal scaling framework is one of the most successful and practically useful frameworks for analyzing the efficiency of MCMC in high-dimensions. It facilitates optimization of the efficiency of one of the most popular class of MCMC algorithms, the Metropolis–Hastings algorithms. The optimal scaling framework provides clear and mathematically based guidance on how to tune the parameters of the Metropolis–Hastings algorithms to asymptotically optimize the performance, and yields guidance for self-tuning or adaptive MCMC methodologies.

To this day, optimal scaling results have mainly been established for target distributions with a product i.i.d. structure, which severely limits their applicability. From a practitioner's perspective, i.i.d. target distributions are too limited to be useful, since they can be tackled by one-dimensional target due to the product structure. On the other hand, practitioners use these tuning criteria far outside the class of i.i.d. target distributions. Extensive simulations show that these optimality results also hold for more complex target distributions. Therefore, there appears to be a gap between theory and practice for tuning Metropolis–Hastings algorithms on general target distributions arising from realistic MCMC models.

In this thesis, we significant narrow the gap between theory and practice of the optimal scaling framework by removing a key limitation. We consider the optimal scaling of random-walk Metropolis algorithms on general target distributions in high dimensions that encompass many realistic MCMC models. We show that asymptotically optimal efficiency can be obtained under general sufficient conditions on the target distribution. The new sufficient conditions are easy to check in practice and hold for some general classes of realistic MCMC models. Our results substantially generalize the product i.i.d. condition. We expect the techniques developed can be applied to other Metropolis–Hastings algorithms as well. Our work also opens up opportunities for developing self-tuning and adaptive methodologies for speeding up the convergence of Metropolis–Hastings algorithms in high dimensions.

1.3 Understanding Bayesian model misspecification

Bayesian analysis implicitly assumes that the statistician has faithfully represented their uncertainty in terms of a statistical model. However, as George Box is often quoted, "all models are wrong". Indeed, in every nontrivial setting, the statistical model is, by necessity, only an approximate representation of uncertainty. In this case, we say that the model is *misspecified*. In practice, the errors introduced by the approximate statistical model can have a serious effect on Bayesian inference and prediction. We need theoretical foundations for Bayesian analysis that do not start by assuming the model is well specified.

One of the hallmarks of Bayesian analysis is the use of "prior" distributions. If we adopt the classical notion of a statistical model, where the data are assumed to be distributed according to one in a parametric family of probability distributions, the prior distribution is a probability measure on the space of parameters that can be used to capture the statistician's prior knowledge and/or subjective beliefs about which parameters are most likely. The combination of the prior and model yields a joint distribution on the data and parameters. If a model is misspecified, it is often the case that no prior distribution yields a joint distribution that accurately represents the statistician's uncertainty. In the misspecified setting, the usual notion of a "subjective" prior may make no sense. Indeed, the statistician may believe that every available parameter setting should be assigned zero prior probability. We focus on this setting of a misspecified model and the question of what the prior distribution should be. Our solution is to view inference in pragmatic terms. Relative to one's beliefs, there is a *surrogate* prior that

is most likely to produce the best answers.

In this thesis, we formalize this notion of pragmatic inference. One of the key ideas is that we frame the problem of choosing a surrogate prior as a Bayesian decision theory task. Our goal is to develop the theoretical and algorithmic foundations for choosing optimal surrogate priors in the misspecified setting. Such a theory would give us a better understanding of the role of prior distributions, and would represent a significant contribution to Bayesian analysis. In addition to providing a way to choose optimal surrogate priors, our framework allows us to evaluate the optimality of existing ad-hoc choices of commonly used priors, and to understand the limitations of Bayesian inference. Furthermore, algorithmic result on efficiently computing (sub-)optimal surrogate priors for complex models would benefit machine learning and other applied fields that use Bayesian analysis. The new foundation of Bayesian model misspecification can be used to seek simple and efficient modifications to traditional Bayesian procedures that are provably robust to misspecification and lead to improved performance.

1.4 Bridging PAC-Bayes theory and Rademacher complexity

Understanding the principles behind the strong empirical performance of modern machine learning algorithms is a problem of great scientific importance. The foundations of machine learning theory are rooted in statistical learning theory, and in particular the decomposition of the risk of learning algorithms into, e.g., approximation and estimation error, and the use of tools from modern probability theory to produce tight estimates of these quantities. Of particular interest are upper bounds on *generalization error*, which is a measure of how accurately an algorithm is able to predict outcome values for previously unseen data.

The PAC-Bayesian theory is one of the most popular tools for developing generalization bounds of learning algorithms. The PAC-Bayes bound is particularly exciting because it provides quantitatively useful results for classifiers with real-valued parameters, which include modern classifiers such as deep neural networks. Recently, there has been a surge of interest and work in PAC-Bayes theory and its application to large-scale neural networks, especially towards studying generalization in overparametrized neural networks trained by variants of gradient descent. PAC-Bayes bounds are one of several tools available for the study of the generalization and risk properties of learning algorithms. Other than PAC-Bayesian bounds, another popular tool for studying generalization is Rademacher complexity, a distribution-dependent complexity measure for classes of real-valued functions.

However, the literature on PAC-Bayes bounds and bounds based on Rademacher complexity are essentially disjoint. PAC-Bayesian theory and Rademacher complexity are widely considered as two different approaches for deriving generalization bounds. The connection between the two is far from clear. In this thesis, we build a bridge between Rademacher complexity and state-of-the-art PAC-Bayesian theory. This is by extending an existing "Rademacher viewpoint" of slow-rate PAC-Bayes bounds to the state-of-art fast-rate PAC-Bayes bounds. We first demonstrate that one can match the fast rate of state-of-art PAC-Bayes bounds using an "extended Rademacher viewpoint" based on shifted Rademacher processes, which is closely related to Rademacher complexity and shifted empirical processes. We then derive a new fast-rate PAC-Bayes bound in terms of the "flatness" of the empirical risk surface, which potentially can be useful for analyzing the generalization of deep neural networks. Our analysis establishes a new framework for deriving fast-rate PAC-Bayes bounds and yields new insights on PAC-Bayesian theory.

1.5 Overview of each chapter

1.5.1 Chapter 2

Markov chain Monte Carlo (MCMC) algorithms [Bro+11; GRS95; Liu08; MT12; RC04] are now routinely used in many fields to obtain approximations of integrals that could not be tackled by common numerical methods, because of the simplicity and the scalability to high-dimensional settings. The running times of MCMC algorithms are an extremely important issue of practice. They have been studied from a variety of perspectives, including convergence "diagnostics" via the Markov chain output [GR92], proving weak convergence limits of sped-up versions of the algorithms to diffusion limits [RGG97; RR98], directly bounding the convergence in total variation distance [MT94; Ros95a; Ros96; RT99; JH01; Ros02; JH04; Bax05; FHJ08], and non-asymptotic guarantees when the target distribution has a smooth and log-concave density, e.g. [BREZ18; Dal17; Dwi+18; DK19] and the references therein.

Among the work of directly bounding the total variation distance, most of the quantitative convergence bounds proceed by establishing a *drift condition* and an associated *minorization condition* for the Markov chain in question (see e.g. [MT12]). One approach for finding quantitative bounds has been the drift and minorization method set forth by Rosenthal [Ros95a].

Computer scientists take a slightly different perspective, in terms of running time complexity order as the "size" of the problem goes to infinity. Complexity results in computer science go back at least to Cobham [Cob65], and took on greater focus with the pioneering NP-complete work of Cook [Coo71]. In the Markov chain context, computer scientists have been bounding convergence times of Markov chain algorithms since at least Sinclair and Jerrum [SJ89], focusing largely on spectral gap bounds for Markov chains on finite state spaces. More recently, attention has turned to bounding spectral gaps of modern Markov chain algorithms on general state spaces, again primarily via spectral gaps, such as [LV03; Vem05; LV06; WSH09a; WSH09b] and the references therein. These bounds often focus on the order of the convergence time in terms of some particular parameter, such as the dimension of the corresponding state space. In recent years, there is much interest in the "large p, large n" or "large p, small n" high-dimensional settings, where p is the number of parameters and n is the sample size. Rajaratnam and Sparks [RS15] use the term convergence complexity to denote the ability of a high-dimensional MCMC scheme to draw samples from the posterior, and how the ability to do so changes as the dimension of the parameter set grows.

Direct total variation bounds for MCMC are sometimes presented in terms of the convergence order, for example, the work by Rosenthal [Ros95b] for a Gibbs sampler for a variance components model. However, current methods for obtaining total variation bounds of such MCMCs typically proceed as if the dimension of the parameter, p, and sample size, n, are fixed. It is thus important to bridge the gap between statistics-style convergence bounds, and computer-science-style complexity results.

In one direction, Roberts and Rosenthal [RR16] connect known results about diffusion limits of MCMC to the computer science notion of algorithm complexity. They show that any weak limit of a Markov process implies a corresponding complexity bound in an appropriate metric. For example, under appropriate assumptions, in p dimensions, the Random-Walk Metropolis algorithm takes $\mathcal{O}(p)$ iterations and the Metropolis-Adjusted Langevin Algorithm takes $\mathcal{O}(p^{1/3})$ iterations to converge to stationarity.

This chapter considers how to obtain MCMC quantitative convergence bounds that can be translated into tight complexity bounds in high-dimensional settings [YR17]. At the first glance, it may seem that an approach to answering the question of convergence complexity may be provided by the drift-and-

minorization method of [Ros95a]. However, Rajaratnam and Sparks [RS15] demonstrate that, somewhat problematically, a few specific upper bounds in the literature obtained by the drift-and-minorization method tend to 1 as n or p tends to infinity. For example, by directly translating the existing work by Choi and Hobert [CH13] and Khare and Hobert [KH13], which are both based on the general approach of [Ros95a], Rajaratnam and Sparks [RS15] show that the "small set" gets large fast as the dimension p increases. And this seems to happen generally when the drift-and-minorization approach is applied to statistical problems. Rajaratnam and Sparks [RS15] also discuss special cases when the method of [Ros95a] can still be used to obtain tight bounds on the convergence rate. However, the conditions proposed in [RS15] are very restrictive. First, it requires the MCMC algorithm to be analyzed is a Gibbs sampler. Second, the Gibbs sampler must have only one high-dimensional parameter which must be drawn in the last step of the Gibbs sampling cycle. Unfortunately, other than some tailored examples [RS15], most realistic MCMC algorithms do not satisfy these conditions. It is unclear whether some particular drift functions lead to bad complexity bounds or the drift-and-minorization approach itself has some limitations. It is therefore the hope by Rajaratnam and Sparks [RS15] that proposals and developments of new ideas analogous to those of [Ros95a], which are suitable for high-dimensional settings, can be motivated.

In this chapter, we attempt to address concerns about obtaining quantitative bounds that can be translated into tight complexity bounds. We note that although Rajaratnam and Sparks [RS15] provide evidence for the claim that many published bounds have poor dependence on n and p, the statistics literature has not focused on controlling the complexity order on n and p. We give some intuition why most directly translated complexity bounds are quite loose and provide advice on how to obtain tight complexity bounds for high-dimensional Markov chains. The key ideas are (1) the drift function should "capture" the posterior modes as n and/or p goes to infinity and (2) "bad" states which have poor drift property when n and/or p gets large should be ruled out when establishing the drift condition. In order to get tight complexity bounds, we propose a modified drift-and-minorization approach by establishing a generalized drift condition for a subset of the state space, which is called the "large set", instead of the whole state space; see Section 2.1. The "large set" is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. By establishing the generalized drift condition, a new quantitative bound is obtained, which is composed of two parts. The first part is an upper bound on the probability the Markov chain will visit the states outside of the "large set"; the second part is an upper bound on the total variation distance of a constructed restricted Markov chain defined only on the "large set". In order to obtain good complexity bounds for high-dimensional settings, the drift function should be chosen to "capture" the posterior modes (this is called a "centered" drift function in [QH17]), and the "large set" should be adjusted depending on n and p to balance the complexity order of the two parts.

As a demonstration, we prove that a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$ iterations. To be more specific, we prove that when the dimension of the model is large, the number of iterations which guarantees small distance of the Gibbs sampler to stationarity is upper bounded by some constant which does not depend on the dimension of the model; see Theorem 2.2.6. As far as we know, this is the first successful example for analyzing the convergence complexity of a *non-trivial* realistic MCMC algorithm using the (modified) drift-and-minorization approach. Several months after we uploaded this manuscript to arXiv, Qin and Hobert [QH17] successfully analyzed another realistic MCMC algorithm using the drift-and-minorization approach. Although the analysis by Qin and Hobert [QH17] does not make use of the "large set" technique proposed in this chapter, they do make use of

a "centered" drift function. We explain in this chapter that when there exists some "bad" states, using a "centered" drift function might not be enough to establish a tight complexity bound. For example, for the Gibbs sampler we analyze in this work, we could not obtain tight complexity bound by the traditional drift-and-minorization approach even if the drift function is "centered". In this sense, even the Gibbs sampler analyzed in this work is a specific example, the analysis of its convergence complexity is technically more difficult than the example in [QH17]. Our modified drift-and-minorization method combining the "large set" technique with "centered" drift function provides a flexible tool for analyzing convergence complexity. It is our hope that this modified drift-and-minorization method of proof in Section 2.1 can be employed to other specific examples for obtaining quantitative bounds that can be translated to complexity bounds in high-dimensional settings.

1.5.2 Chapter 3

The optimal scaling framework [RGG97; RR98; RR01] is one of the most successful and practically useful ways of performing asymptotic analysis of MCMC methods in high-dimensions. Optimal scaling results (e.g. [CRR05; NR06; Béd08; BR08; NR08; NR11; NRY12; JLM15; JLM14; RR14; ZBK17]) facilitate optimization of MCMC performance by providing clear and mathematically-based guidance on how to tune the parameters defining the proposal distribution for Metropolis–Hastings algorithms [Met+53; Has70]. For instance, classical results include tuning the acceptance probabilities to 0.234 for random-walk Metropolis algorithm (RWM) [RGG97] and 0.574 for Metropolis-adjusted Langevin algorithm (MALA) [RR98]. Moreover, optimal scaling results have been used to analyze and compare a wide variety of MCMC algorithms, such as Hamiltonian Monte Carlo (HMC) [Bes+13], Pseudo-Marginal MCMC [She+15], multiple-try MCMC [BDM12]. This yields guidance which is widely used by practitioners, especially via self-tuning or Adaptive MCMC methodologies [AT08; Ros11].

In the original paper, Roberts, Gelman, and Gilks [RGG97] dealt with the RWM algorithm starting in stationarity for target distributions which have i.i.d. product forms. The i.i.d. condition for the target and the assumption for the chain to start in stationarity are two main limitations of the optimal scaling framework. Particularly, the product i.i.d. condition is very restrictive. From a practitioner's perspective, target distributions of the i.i.d. forms are too limited a class of probability distributions to be useful, since they can be tackled by sampling a single one-dimensional target due to the product structure. To this day, optimal scaling results have mainly been proved for target distributions with a product structure, which severely limits their applicability. On the other hand, practitioners use these tuning criteria far outside the class of target distributions of product i.i.d. forms. For example, extensive simulations [RR01; SFR10] show that these optimality results also hold for more complex target distributions.

There exists only a few extensions for correlated targets and most of them are derived for very specific models. For example, Breyer and Roberts [BR00] studied target densities which are Gibbs measures and Roberts and Rosenthal [RR01] studied inhomogeneous target densities. Breyer, Piccioni, and Scarlatti [BPS04] studied target distributions arising in nonlinear regression and have a mean field structure. Neal and Roberts [NR06] considered the case where updates of high-dimensional Metropolis algorithms are lower dimensional than the target density itself. Later, Bédard and Rosenthal [BR08] studied independent targets with different scales (see also [Béd07; Béd08]) and Bédard [Béd19] studied a special family of hierarchical target distributions. Neal and Roberts [NR08] studied spherically constrained target distributions and non-Gaussian proposals [NR11]. Sherlock and Roberts [SR09] considered elliptically symmetric unimodal targets. Neal, Roberts, and Yuen [NRY12] studied densities with bounded support.

Durmus et al. [Dur+17] considered target distributions which are differentiable in L^p mean. Recently, Mattingly, Pillai, and Stuart [MPS12] studied diffusion limits for a class of high-dimensional measures found from the approximation of measures on a Hilbert space which are absolutely continuous with respect to a Gaussian reference measure (See also [PST12; BRS09; Bes+08; Cot+13]). Important examples of this scenario required by [MPS12] in uncertainty quantification problems are given in [HMS11; Stu10; Che+18]. However, in this chapter we shall concentrate on the situation where absolute continuity with respect to a Gaussian is not a reasonable assumption, as is the case in many Bayesian statistics applications.

Furthermore, we do not consider the transient phase of the Metropolis–Hasting algorithms in this work. The transient phase of high-dimensional Metropolis–Hasting algorithms are studied for example in [CRR05; JLM14; JLM15; KOS18; KOS19]. Kuntz, Ottobre, and Stuart [KOS19] studied the RWM algorithm starting out of stationarity in the settings of [MPS12; JLM15] when non-product target distributions are defined in a Hilbert space being absolute continuous with respect to some Gaussian measures. Such target distributions in [KOS19] can arise for example in Bayesian nonparametric settings, but not in many other Bayesian statistics applications which we focus on in this work.

In this chapter, we consider optimal scaling of RWM algorithms on general target distributions in high dimensions arising from realistic MCMC models [YRR19]. First, for optimal scaling by maximizing expected squared jumping distance (ESJD), we show the asymptotically optimal acceptance rate 0.234 can be obtained under general sufficient conditions on the target distribution. Very briefly speaking, 0.234 is asymptotically optimal if (i) each coordinate of the Markov chain is only strongly dependent with a subset of other coordinates (see assumptions A1 and A3); (ii) the target distribution satisfies some smoothness conditions (see assumptions A2 and A4); (iii) as the dimension goes to infinity, a key quantity of "roughness" of the target concentrates to a nonzero value (see assumption A5). The new sufficient conditions are easy to check in practice and may hold for some general classes of realistic MCMC models. Our results substantially generalize the commonly used product i.i.d. condition. Furthermore, we show one-dimensional diffusion limits can also be obtained under relaxed conditions which still allow dependent coordinates of the target distribution. Finally, we also connect the new results of diffusion limits to complexity bounds of RMW algorithms in high dimensions.

The chapter is organized as follows. In Section 3.1, we give a brief background review of optimal scaling for Metropolis–Hastings algorithms and complexity bounds via diffusion limits. In Section 3.2, we present our main results, which include three parts: optimal scaling by maximizing ESJD, optimal scaling via diffusion limits, and complexity bounds via diffusion limits. In Section 3.3, we demonstrate the new optimal scaling result holds for some realistic MCMC models. In Section 3.4.1, we prove Theorem 3.2.10, which is one of our main results. The proofs of lemmas used for proving Theorem 3.2.10 and other main results, such as Theorems 3.2.19 and 3.2.21, are delayed to Sections 3.4.2, 3.4.3 and 3.4.5.

1.5.3 Chapter 4

Faced with a decision under uncertainty, where each course of action $\alpha \in \mathcal{A}$ incurs an unknown loss $\bar{L}(\alpha)$, the Bayesian decision-theoretic approach treats the unknown losses $\bar{L}(\alpha)$, $\alpha \in \mathcal{A}$, as random variables, whose joint distribution represents the statistician's subjective uncertainty in mathematical form. In light of additional evidence, modeled by additional random variables defined on the same probability space, the Bayesian chooses an action α that minimizes the conditional expectation of $\bar{L}(\alpha)$.

In practice, the Bayesian's task of representing their uncertainty in terms of random variables modeling

losses and evidence is formidable. (We will refer to a probability distribution over losses and evidence as a belief distribution.) Even given a belief distribution, one often faces an intractable computation trying to identify the action that minimizes the conditional expected loss. Approximations at each stage can have serious real-world consequences. Closing this gap between the theory and practice of Bayesian decision theory is the single greatest problem facing its development.

In this work, we focus on the problem of *Bayesian misspecification*, where one cannot produce a belief distribution that matches one's subjective uncertainty. There is a large literature on this problem with a host of prescriptions [Ber66; Che85; BM98; GPW01; Wal02; RT03; Mül13; MD15; PV17].

To address the challenge of Bayesian misspecification, we take the stance that the Bayesian's belief distribution is, in practice, *chosen* among alternatives, rather than determined. We then cast this *decision* back into the framework of Bayesian decision theory. The resulting framework—which we call *meta-Bayesianism*—is the study of *pairs* of belief distributions, one presumed to describe our actual subjective beliefs, and the other misspecified or thrust upon us, yet ultimately constraining the decisions we will make in light of evidence. This perspective on belief distributions can be seen as a formalization of pragmatism: the *optimal* belief distribution is that which leads to decisions that minimize our expected loss, relative to our actual subjective beliefs. Put simply, an optimal belief distribution leads us to make the best possible decisions, even if it does not represent our subjective beliefs accurately. Crucially, this perspective allows for nonoptimality: the choice to cut short or approximate the process of forming a belief distribution can be justified on the grounds that one believes the gap is unlikely to have a large impact on our loss.

Performing a meta-Bayesian analysis is, in general, no easier than performing a Bayesian analysis with respect to one's actual subjective beliefs. Therefore, the central goal of meta-Bayesian decision theory is to identify broadly applicable theorems that guide the practical development of belief distributions. Such results would give us a better understanding of the role of belief distributions in statistical practice, and would represent a significant contribution to the theoretical foundations of Bayesian decision theory.

Ultimately, we may not only be concerned with expected loss, but also with the computational resources consumed in the process of decision making. These and other criteria can be incorporated into the meta-Bayesian framework. In addition to providing a way to choose optimal belief distributions, our framework allows us to evaluate the optimality of existing ad-hoc choices of commonly used priors, and to understand the limitations of Bayesian inference under misspecification.

In this chapter, we present early results in meta-Bayesian analysis for the problem of choosing an optimal prior for a misspecified model. In Section 4.2, we first introduce some preliminary results, including Bayesian decision theory in a more general setting. In Section 4.3.1, we give a formal definition of meta-Bayesian decision principle and in Section 4.3.2, we introduce a particular setting for meta-Bayesian analysis for prediction tasks. In Section 4.3.3, we consider meta-Bayesian decision problem for choosing prior in the prediction setting. We demonstrate how the loss function under consideration affects the definition of the optimal surrogate prior. For example, under the classic self-information loss (i.e., log loss), the optimal surrogate prior minimizes the conditional relative entropy [Gra90], while, under the quadratic loss, the optimal surrogate prior minimizes the expected quadratic distance between the true and model posterior means. We also discuss sufficient conditions for a surrogate prior to be optimal. In Section 4.3.4, we consider general cases when the belief is a mixture of i.i.d. distributions. We show that, under certain conditions, the belief on the asymptotic "locations" of the posterior distributions is indeed an asymptotically optimal prior, which supports the heuristics from the existing results by

[Ber66] on asymptotic behavior of posterior distributions. Finally, in Section 4.3.5 and Section 4.3.6, we study a specific example in the setting of the i.i.d. Bernoulli model versus more general beliefs on binary sequence. For example, for the case of i.i.d. Bernoulli sequences, under the log loss, the optimal surrogate prior only depends on the first $n + k$ moments of the prior belief, where n denotes the number of data points you plan to observe and k denotes the number of predictions you expect to make. We also prove that, for the i.i.d Bernoulli model versus any belief on binary sequence, under some conditions, the marginal prior belief on the limiting frequency of ones is asymptotically optimal when the number of predictions you expect to make goes to infinity.

1.5.4 Chapter 5

PAC-Bayes theory [McA99; STW97] was developed to provide *probably approximately correct* (PAC) guarantees for supervised learning algorithms whose outputs can be expressed as a weighted majority vote. Its uses have expanded considerably since [LLST13; AB07; Bég+16; Ger+16; Thi+16; GM19; SL17; Gue19]. See [Lan05; Erv14; McA13] for gentle introductions. Indeed, there has been a surge of interest and work in PAC-Bayes theory and its application to large-scale neural networks, especially towards studying generalization in overparametrized neural networks trained by variants of gradient descent [DR17; DR18a; DR18b; NBS17; Ney+17; Lon17].

PAC-Bayes bounds are one of several tools available for the study of the generalization and risk properties of learning algorithms. One advantage of the PAC-Bayes framework is its ease of use: one can obtain high-probability risk bounds for arbitrary ("posterior") Gibbs classifiers provided one can compute or bound relative entropies with respect to some fixed ("prior") Gibbs classifier. Another tool for studying generalization is Rademacher complexity, a distribution-dependent complexity measure for classes of real-valued functions [BM02; KP02; BBM05; LRS15; Men14; ZH18].

The literature on PAC-Bayes bounds and bounds based on Rademacher complexity are essentially disjoint. One point of contact is the work of Kakade, Sridharan, and Tewari [KST08], which builds the first bridge between PAC-Bayes theory and Rademacher complexity. By viewing Gibbs classifiers as linear operators and relative entropy as a strictly convex regularizer, they were able to use their general Rademacher complexity bounds on strictly convex linear classes to develop a slightly sharper version of McAllester's PAC-Bayes bound [McA99]. This result offers new insight on PAC-Bayes theory, including potential roles for data-dependent complexity estimates and stability. However, even within the PAC-Bayes community, this result is relatively unknown.

While the PAC-Bayes bound established by Kakade, Sridharan, and Tewari improves on McAllester's bound, it still converges at a slow $1/\sqrt{m}$ rate, where m denotes the number of data used to form the empirical risk estimate. This observation raises the question of whether one can match state-of-the-art PAC-Bayes bounds via a Rademacher-process argument. In particular, can one match Catoni's bound [Cat07, Thm. 1.2.6], which can obtain a fast $1/m$ rate of convergence?

There is an extensive literature on the problem of obtaining fast $1/m$ rates of convergence for the generalization error of (approximate) empirical risk minimization (ERM). Available approaches include the use of local Rademacher complexity [BBM05; Kol06], shifted empirical processes [LM12], offset Rademacher complexities [LRS15], and local empirical entropy [ZH18]. See also [MN06; GK06; HY15; Han16; LM13; Men17] and [van+15] for an extensive survey. To date, these techniques have not been connected to PAC-Bayesian theory, which presents the opportunity to obtain new PAC-Bayes theory for ERM.

In this chapter, we extend the bridge between Rademacher process theory and PAC-Bayes theory by constructing new bounds using Rademacher process techniques [YSR19]. Among our contributions:

1. We show how to recover Catoni's fast-rate PAC-Bayes bound [Cat07], up to constants, using tail bounds on shifted Rademacher processes, which are special cases of shifted empirical processes [Weg03; LM12; ZH18]; See Section 5.2.

2. We derive a new fast-rate PAC-Bayes bound, building on our shifted-Rademacher-process approach. This bound is determined by the "flatness" of the empirical risk surface on which the posterior Gibbs classifier concentrates. The notion of "flatness" is inspired by the proposal by Dziugaite and Roy [DR17] to formalize the empirical connection between "flat minima" and generalization using PAC-Bayes bounds; See Section 5.3.

3. More generally, we introduce a new approach to derive fast-rate PAC-Bayes bounds and, in turn, offer new insight on PAC-Bayesian theory.

Chapter 2

Complexity results for MCMC derived from quantitative bounds

This chapter considers how to obtain MCMC quantitative convergence bounds which can be translated into tight complexity bounds in high-dimensional settings. We propose a modified drift-and-minorization approach, which establishes a generalized drift condition defined in a subset of the state space. The subset is called the "large set", and is chosen to rule out some "bad" states which have poor drift property when the dimension gets large. Using the "large set" together with a "centered" drift function, a quantitative bound can be obtained which can be translated into a tight complexity bound. As a demonstration, we analyze a certain realistic Gibbs sampler algorithm and obtain a complexity upper bound for the mixing time, which shows that the number of iterations required for the Gibbs sampler to converge is constant under certain conditions on the observed data and the initial state. It is our hope that this modified drift-and-minorization approach can be employed in many other specific examples to obtain complexity bounds for high-dimensional Markov chains.

2.1 Generalized geometric drift conditions and large sets

We use $\xrightarrow{d}$ for weak convergence and $\pi(\cdot)$ to denote the stationary distribution of the Markov chain. The total variation distance is denoted by $\|\cdot\|_{\text{var}}$ and the law of a random variable X denoted by $\mathcal{L}(X)$. We adopt the Big-O, Little-O, Theta, and Omega notations. Formally, $T(n) = \mathcal{O}(f(n))$ if and only if for some constants c and n_0, $T(n) \leq cf(n)$ for all $n \geq n_0$; $T(n) = \Omega(f(n))$ if and only if for some constants c and n_0, $T(n) \geq cf(n)$ for all $n \geq n_0$; $T(n)$ is $\Theta(f(n))$ if and only if both $T(n) = \mathcal{O}(f(n))$ and $T(n) = \Omega(f(n))$; $T(n) = o(f(n))$ if and only if $T(n) = \mathcal{O}(f(n))$ and $T(n) \neq \Omega(f(n))$.

Scaling classical MCMCs to very high dimensions can be problematic. Even if a chain is geometrically ergodic for fixed n and p, the convergence of Markov chains may still be quite slow as $p \to \infty$ and $n \to \infty$. For a Markov chain $\{X^{(i)}, i = 0, 1, \dots\}$ on a state space $(\mathcal{X}, \mathcal{B})$ with transition kernel $P(x, \cdot)$, defined by

$$P(x, B) = \mathbb{P}(X^{(i+1)} \in B \,|\, X^{(i)} = x), \quad \forall x \in \mathcal{X}, B \in \mathcal{B} \tag{2.1}$$

the general method of [Ros95a] proceeds by establishing a *drift condition*

$$\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x) \leq \lambda f(x) + b, \quad \forall x \in \mathcal{X}, \tag{2.2}$$

where $f : \mathcal{X} \to \mathbb{R}^+$ is the "drift function", some $0 < \lambda < 1$ and $b < \infty$; and an associated *minorization condition*

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{2.3}$$

where $R := \{x \in \mathcal{X} : f(x) \leq d\}$ is called the "small set", and $d > 2b/(1-\lambda)$, for some $\epsilon > 0$ and some probability measure $Q(\cdot)$ on $\mathcal{X}$. Then [Ros95a, Theorem 12] states that under both drift and minorization conditions, if the Markov chain starts from an initial distribution ν, then for any $0 < r < 1$, we have

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1-\epsilon)^{rk} + \alpha^{-k}(\alpha\Lambda)^{rk}\left[1 + \mathbb{E}_\nu(f(x)) + \frac{b}{1-\lambda}\right], \tag{2.4}$$

where $\alpha^{-1} = \frac{1+2b+\lambda d}{1+d}$, $\Lambda = 1 + 2(\lambda d + b)$ and $\mathbb{E}_\nu[f(x)]$ denotes the expectation of $f(x)$ over $x \sim \nu(\cdot)$. However, it is observed, for example, in [RS15; QH17], that for many specific bounds obtained by the drift-and-minorization method, when the dimension gets larger, the typical scenario for the drift condition of Eq. (2.2) seems to be λ going to one, and/or b getting much larger. This makes the "size" of the small set R grow too fast, which leads to the minorization volume ϵ go to 0 exponentially fast. In the following, we give an intuitive explanation of what makes a "good" drift condition in high-dimensional settings.

2.1.1 Intuition

It is useful to think of the drift function $f(x)$ as an energy function [JH01]. Then the drift condition in Eq. (2.2) implies the chain tends to "drift" toward states which have "lower energy" in expectation. It is well-known that a "good" drift condition is established when both λ and b are small. Intuitively, λ being small implies that when the chain is in a "high-energy" state, then it tends to "drift" back to "low-energy" states fast; and b being small implies that when the chain is in a "low-energy" state, then it tends to remain in a "low-energy" state in the next iteration too. In a high-dimensional setting as the dimension grows to infinity, for a collection of drift conditions to be "good", we would like it to satisfy the following two properties:

P1. λ is small, in the sense that it converges to 1 slowly or is bounded away from 1;

P2. b is small, in the sense that it grows at a slower rate than do typical values of the drift function.

One way to understand this intuition is to think of it as controlling the complexity order of the size of the "small set", $R = \{x \in \mathcal{X} : f(x) \leq d\}$. Since $d > 2b/(1-\lambda)$, if λ converges to 1 slowly or is bounded away from 1, and if b is growing at a slower rate than typical values of $f(x)$, then the size of the small set parameter d can be chosen to have a small complexity order on n and/or p. This in turn makes the minorization volume ϵ converge to 0 sufficiently slowly (or even remain bounded away from 0).

Next, we provide some advice on how to establish such a "good" drift condition in high-dimensional settings.

For clarity, we first assume that λ is bounded away from 1, and focus on conditions required for b to grow at a slower rate than typical values of $f(x)$. Assume for definiteness that p is fixed and $n \to \infty$,

and the drift function is scaled in such a way that $f(x) = \mathcal{O}(1)$ and there is a fixed typical state $\tilde{x}$ with $f(\tilde{x}) = \Theta(1)$ regardless of dimension. Then, to satisfy property P2 above, we require that $b = o(1)$. On the other hand, taking expectation over $x \sim \pi(\cdot)$ on both sides of Eq. (2.2) yields $b \geq \mathbb{E}_\pi[f(x)]/(1-\lambda)$, so $b = \Omega(\mathbb{E}_\pi[f(x)])$. To make $b = o(1)$ implies that the drift function should be chosen such that

$$\mathbb{E}_\pi[f(x)] \to 0.$$

Therefore, to get a small b in a high-dimensional setting, we require a (properly scaled) drift function $f(\cdot)$ whose values $f(x)$, where $x \sim \pi(\cdot)$, concentrate around 0. In particular, if the stationary distribution $\pi(\cdot)$ concentrates near multiple modes as $n \to \infty$, then to make $\mathbb{E}_\pi[f(x)] \to 0$, we require a drift function which "captures" the modes in the sense of nearly vanishing near them. In this work, we use the name "centered" drift functions [QH17] to denote drift functions that "capture" the modes of the stationary distribution $\pi(\cdot)$ in this sense.

Note that in the literature, the drift functions used to establish the drift condition are usually not "centered". This is because in the traditional setting where n and p are fixed, a "good" drift condition is established whenever λ and b are small enough for specific fixed values of n and p. The complexity orders of λ and b as functions of n and/or p are not essential, so the property of "capturing" the posterior modes is not necessary for establishing a good drift condition. As a result, many existing quantitative bounds cannot be directly translated into tight complexity bounds, since the size of the small set does not have a small complexity order on n and/or p. At the very least, one has to re-analyze such MCMC algorithms using "centered" drift functions.

Next, we focus on establishing λ that is either bounded away from 1 or converges to 1 slowly, assuming the drift function is already chosen to be "centered". Intuitively, λ describes the behavior of the Markov chain when its current state has a "high energy". If λ goes to 1 very fast when n and/or p goes to infinity, this may suggest the existence of some "bad" states, i.e. states which have "high energy", but the drift property becomes poor as n and/or p gets large. Therefore, in high dimensions, once the Markov chain visits in one of these "bad" states, it only slowly drifts back toward to the corresponding small set. Since the drift condition in Eq. (2.2) must hold for all $x \in \mathcal{X}$, the existence of "bad" states forces λ go to 1 very fast. And since the small set is defined as $R = \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda)$, the scenario $\lambda \to 1$ very fast forces R to become very large, and hence the minorization volume ϵ goes to zero very fast. One perspective on this problem is that the definition of drift condition in Eq. (2.2) is too restrictive, since it must hold for all states x, even the bad ones.

In summary, we are able to establish a small b as in P2 above by simply using a "centered" drift function. However, the main difficulty in establishing a small λ as in P1 above is the existence of some "bad" states when n and/or p gets large. Since the traditional drift condition defined in Eq. (2.2) is restrictive, the traditional drift-and-minorization method is not flexible enough to deal with these "bad" states. In this following, we instead propose a modified drift-and-minorization approach using a generalized drift condition, where the drift function is defined only in a "large set". This allows us to rule out those "bad" states in high-dimensional cases.

2.1.2 New quantitative bound

We first relax the traditional drift condition and define a generalized drift condition which is established only on a subset of the state space. Let $\{X^{(k)}\}$ be a irreducible Markov chain on a state space $(\mathcal{X}, \mathcal{B})$

with a transition kernel $P(x,\cdot), \forall x \in \mathcal{X}$, $P^k(x,\cdot)$ be the k-step transition kernel and π be the stationary distribution of the Markov chain. Denote R_0 as the "large set", i.e., $R_0 \in \mathcal{B}$ is a subset of $\mathcal{X}$.

Definition 2.1.1. (Generalized drift condition on a large set) There exists a drift function $f : \mathcal{X} \to \mathbb{R}^+$ such that for some $\lambda < 1$ and $b < \infty$,

$$\mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x) \leq \lambda f(x) + b, \quad \forall x \in R_0, \tag{2.5}$$

and (C1) *or* (C1') holds.

(C1). The "large set" R_0 is defined by $R_0 = \{x \in \mathcal{X} : f(x) \leq d_0\}$ for some $d_0 > 0$.

(C1'). The transition kernel $P(x,\cdot)$ can be written as a composition of reversible steps $P = \prod_{i=1}^{I} P_i$, i.e. , $P(x, \mathrm{d}y) = \int_{(x_1,\ldots,x_{I-1})\in\mathcal{X}\times\cdots\times\mathcal{X}} P_1(x,\mathrm{d}x_1)P_2(x_1,\mathrm{d}x_2)\cdots P_I(x_{I-1},\mathrm{d}y)$, where $I \geq 1$ is a fixed integer, and

$$\mathbb{E}(f(\tilde{X}^{(1)}) \,|\, \tilde{X}^{(0)} = x) \leq \mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x), \quad \forall x \in R_0, \tag{2.6}$$

where $\{\tilde{X}^{(k)}\}$ denotes a restricted Markov chain with a transition kernel $\prod_{i=1}^{I} \tilde{P}_i$ where $\tilde{P}_i(x,\mathrm{d}y) := P_i(x,\mathrm{d}y)$ for $x, y \in R_0, x \neq y$, and $\tilde{P}_i(x,x) := 1 - P_i(x, R_0\backslash\{x\}), \forall x \in R_0$.

Remark 2.1.2. Note that only one of (C1) and (C1') is required. For (C1'), the Markov chain needs to be either reversible or can be written as a composition of reversible steps. This condition is very mild since it is satisfied by most realistic MCMC algorithms. For example, full-dimensional and random-scan Metropolis-Hastings algorithms and random-scan Gibbs samplers are reversible, and their deterministic-scan versions can be written as a composition of reversible steps. For (C1), it is required that the "large set" is constructed using the drift function in a certain way but there is no restriction for the transition kernel P. If R_0 is constructed as in (C1) then Eq. (2.6) automatically holds. Therefore, one should verify (C1') if one hopes to have more flexibility for constructing R_0 than the particular way in (C1). Particularly, if the drift function $f(x)$ depends on all coordinates, it might be hard to control all the states in $\{x \in \mathcal{X} : f(x) \leq d_0\}$ as the dimension increases. Then (C1') might be preferable.

Remark 2.1.3. To verify (C1') in Definition 2.1.1, one has to check a new inequality $\mathbb{E}(f(\tilde{X}^{(1)}) \,|\, \tilde{X}^{(0)} = x) \leq \mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x)$. This inequality in (C1') implies the "large set" R_0 should be chosen such that the states in R_0 have "lower energy" on expectation. This is intuitive since we assume the "bad" states all have "high energy" and poor drift property when n and/or p gets large. One trick is to choose R_0 by ruling out some (but not too many) states with "high energy" even if the states are not "bad". In Section 2.2, we demonstrate the use of this trick to select the "large set" R_0 so that $\mathbb{E}(f(\tilde{X}^{(1)}) \,|\, \tilde{X}^{(0)} = x) \leq \mathbb{E}(f(X^{(1)}) \,|\, X^{(0)} = x)$ can be easily verified. The constructed R_0 in Section 2.2 satisfies (C1') but not (C1).

Next, we propose a new quantitative bound, which is based on the generalized drift condition on a "large set".

Theorem 2.1.4. *Suppose the Markov chain satisfies the generalized drift condition in Definition 2.1.1 on a "large set" R_0. Furthermore, for a "small set" $R := \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda)$, the Markov chain also satisfies a minorization condition:*

$$P(x,\cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{2.7}$$

for some $\epsilon > 0$, some probability measure $Q(\cdot)$ on $\mathcal{X}$. Finally, suppose the Markov chain begins with an initial distribution ν such that $\nu(R_0) = 1$. Then for any $0 < r < 1$, we have

$$\begin{aligned}\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq (1 - \epsilon Q(R_0))^{rk} + \frac{(\alpha\Lambda)^{rk}\left[1 + \mathbb{E}_\nu[f(x)] + \frac{b}{1-\lambda}\right] - \alpha^{rk}}{\alpha^k - \alpha^{rk}} \\ + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c),\end{aligned} \tag{2.8}$$

where $\alpha^{-1} = \frac{1+2b+\lambda d}{1+d}$, $\Lambda = 1 + 2(\lambda d + b)$.

Proof. See Section 2.3.1. □

Remark 2.1.5. Note that the new bound in Theorem 2.1.4 assumes the Markov chain begins with an initial distribution ν such that $\nu(R_0) = 1$. This assumption is not very restrictive since the "large set" ideally should include all "good" states. In high-dimensional settings, the Markov chain is not expected to converge fast beginning with any state (see Section 2.2.2 for discussions on initial states). For the term $Q(R_0)$ in Eq. (2.8), it can be replaced by any lower bound of $Q(R_0)$. Since the "large set" is ideally chosen to include all "good" states, one can expect $Q(R_0)$ is at least bounded away from 0. In particular, if we have established an upper bound for $P(x, R_0^c)$ with $x \in R$, then we can apply $\epsilon Q(R_0^c) \leq P(x, R_0^c)$ to get an upper bound of $Q(R_0^c)$ which can be turned into a lower bound on $Q(R_0)$.

Remark 2.1.6. In the proof of Theorem 2.1.4, the generalized drift condition in Definition 2.1.1 essentially implies a traditional drift condition in Eq. (2.2) for a constructed "restricted" Markov chain only on the "large set" R_0. The first two terms in the upper bound Eq. (2.8) are indeed an upper bound on the total variation distance of this constructed "restricted" Markov chain. Note that the general idea of studying the restriction of a Markov chain to some "good" subset of the state space has appeared in the literature, such as [MR00; DF03; Jer+04; Eft+16; MS17; RS18; MARS19] and the references therein, in which different ways of restrictions have been considered for different reasons. For example, Bou-Rabee and Hairer [BRH13] studied the rate of convergence of the MALA algorithm by a similar argument, which is later extended in [Ebe14] to study contraction rate in Wasserstein distance w.r.t. Gaussian reference measure. However, the argument in [BRH13] is only for the MALA algorithm and the proof technique is by constructing a restricted chain. Comparing with [BRH13], our Theorem 2.1.4 is for general MCMC algorithms with weaker conditions in (C1) and (C1'). In the proof, we use either a trace chain or a restricted chain depending on which condition is satisfied. Most importantly, the motivation of this work is to obtain tight complexity bound which is quite different from [BRH13]. In Theorem 2.1.4, the goal of considering a "good" subset of the state space is to obtain better control on the dependence on n and p for the upper bound.

Remark 2.1.7. The last two terms in the upper bound Eq. (2.8) give an upper bound of the probability that the Markov chain will visit R_0^c starting from either the initial distribution ν or the stationary distribution π. Therefore, the proposed method in Theorem 2.1.4 is a generalized version of the classic drift-and-minorization method [Ros95a] by allowing the drift condition is established on a chosen "large set". Indeed, if we choose $R_0 = \mathcal{X}$, then Eq. (2.8) is almost the same as Eq. (2.4), except slightly tighter due to the terms α^{rk}.

Remark 2.1.8. One more note about Eq. (2.8) is that the new bound does not decrease exponentially with k. For example, the term $k\,\pi(R_0^c)$ is linear increasing with k for fixed n and p. We emphasize that

we do not aim to prove a Markov chain is geometrically ergodic here. An upper bound which decreases exponentially with k for fixed n and p does not guarantee to have a tight complexity order on n and/or p, which has been discussed in [RS15]. Instead, our new bound in Eq. (2.8) is designed for controlling complexity orders of n and/or p for high-dimensional Markov chains.

Remark 2.1.9. The Markov chain to be analyzed in Theorem 2.1.4 does not have to be geometrically ergodic. The proof of Eq. (2.8) only implies that, after ruling out "bad" states, a constructed "restricted" Markov chain defined on the "large set" is geometrically ergodic. Therefore, technically speaking, the new bound in Eq. (2.8) can be used to analyze non-geometrically ergodic high-dimensional Markov chains.

2.1.3 Complexity bound

The proposed new bound in Theorem 2.1.4 can be used to obtain complexity bounds in high-dimensional settings. The key is to balance the complexity orders of k on n and/or p required for both the first two terms and the last two terms of the upper bound in Eq. (2.8) to be small. The complexity order of k on n and/or p for the first two terms to be small can be controlled by adjusting the "large set". The "large set" should be kept as large as possible provided that "bad" states have been ruled out. For the last two terms to be small, we should determine the growth rate of k as a function of n and p so that

$$k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c) \to 0. \tag{2.9}$$

This may involve (carefully) bounding the tail probability of the transition kernel, depending on the definition of the "large set" and the complexity order aimed to establish.

In the next section, we employ the modified drift-and-minorization method to prove a certain realistic Gibbs sampler algorithm converges in $\mathcal{O}(1)$. We first choose a particular "centered" drift function $f(x)$ and identify the "bad" states. In our Gibbs sampler example, one coordinate of the state x corresponds to one particular parameter of the MCMC model, and the "bad" states correspond to those whose value of this particular parameter is close to zero. Then we define the "large set" by ruling out the "bad" states. This allow us to obtain a quantitative bound using Theorem 2.1.4. Finally, under high-dimensional settings, the obtained quantitative bound can be translated into a complexity bound, which shows that the mixing time of the Gibbs sampler is $\mathcal{O}(1)$. Note that mixing time is often defined uniformly over initial states, which is difficult to extend to general state spaces. In this work, the term "mixing time" is defined depending on the initial state. The formal definition is given in the following.

Definition 2.1.10. For any $0 < c < 1$, we define the mixing time K_c of a Markov chain $\{X^{(k)}\}$ with initial state x by

$$K_c := \arg\min_k \left\{ \|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \le c \right\} \quad \text{such that } X^{(0)} = x. \tag{2.10}$$

2.2 Gibbs sampler convergence bound

We concentrate on a particular MCMC model, which is related to the James-Stein estimators [Ros96]:

$$\begin{aligned} Y_i \mid \theta_i &\sim \mathcal{N}(\theta_i, V), \quad 1 \le i \le n, \\ \theta_i \mid \mu, A &\sim \mathcal{N}(\mu, A), \quad 1 \le i \le n, \\ \mu &\sim \text{ flat prior on } \mathbb{R}, \\ A &\sim \mathbf{IG}(a, b), \end{aligned} \tag{2.11}$$

where V is assumed to be known, $(Y_1, \dots, Y_n)$ is the observed data, and $x = (A, \mu, \theta_1, \dots, \theta_n)$ are parameters. Note that we have the number of parameters $p = n + 2$ in this example. For simplicity, we will not mention p but only refer to n for this model. The posterior distribution satisfies

$$\begin{aligned} \pi(\cdot) &= \mathcal{L}(A, \mu, \theta_1, \dots, \theta_n \mid Y_1, \dots, Y_n) \\ &\propto \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\theta_i - \mu)^2}{2A}} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(Y_i - \theta_i)^2}{2V}}. \end{aligned} \tag{2.12}$$

A Gibbs sampler for the posterior distribution of this model has been originally analyzed in [Ros96]. A quantitative bound has been derived by Rosenthal [Ros96] using the drift-and-minorization method with a drift function $f(x) = \sum_{i=1}^{n} (\theta_i - \bar{Y})^2$ where $\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$. We first observe that this drift function is not "centered". For example, select a "typical" state $\tilde{x} = (\tilde{A}, \tilde{\mu}, \tilde{\theta}_1, \dots, \tilde{\theta}_n)$ such that $\tilde{\theta}_i = Y_i$, we get $f(\tilde{x}) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$. Under reasonable assumptions on the observed data $\{Y_i\}$, we can get the properly scaled drift function $\frac{1}{n} f(\tilde{x}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \Theta(1)$. Then if the drift function is "centered", we hope the established b satisfies $b/n = o(1)$. However, $b/n = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + \frac{n+1/4}{n} V = \Theta(1)$ in [Ros96]. Furthermore, the established λ in [Ros96] converges to 1 very fast, satisfying $1/(1-\lambda) = \Omega(n)$. Therefore, if we translate the quantitative bound in [Ros96] into complexity orders, it requires the size of the "small set" R be $\Omega(n^2)$, which makes the minorization volume ϵ be exponentially small. This leads to upper bounds on the distance to stationarity which require exponentially large number of iterations to become small. This result also coincides with the observations by Rajaratnam and Sparks [RS15] when translating the work of Khare and Hobert [KH13] and Choi and Hobert [CH13].

We demonstrate the use of the modified drift-and-minorization approach by analyzing a Gibbs sampler for this MCMC model. Defining $x^{(k)} = (A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \dots, \theta_n^{(k)})$ to be the state of the Markov chain at the k-th iteration, we consider the following order of Gibbs sampling for computing the posterior distribution:

$$\begin{aligned} \mu^{(k+1)} &\sim \mathcal{N}\left(\bar{\theta}^{(k)}, \frac{A^{(k)}}{n}\right), \\ \theta_i^{(k+1)} &\sim \mathcal{N}\left(\frac{\mu^{(k+1)} V + Y_i A^{(k)}}{V + A^{(k)}}, \frac{A^{(k)} V}{V + A^{(k)}}\right), \quad i = 1, \dots, n, \\ A^{(k+1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2} \sum_{i=1}^{n} (\theta_i^{(k+1)} - \bar{\theta}^{(k+1)})^2\right). \end{aligned} \tag{2.13}$$

We prove that convergence of this Gibbs sampler is actually very fast: the number of iterations required is $\mathcal{O}(1)$. More precisely, we first make the following assumptions on the observed data $\{Y_i\}$: there exists $\delta > 0$, $\bar{V} < \infty$, and a positive integer N_0, such that, almost surely with respect to the randomness of

$\{Y_i\}$:

$$V + \delta \leq \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} \leq \bar{V}, \quad \forall n \geq N_0. \tag{2.14}$$

Remark 2.2.1. The assumption in Eq. (2.14) is quite natural. The upper bound is just to ensure $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \mathcal{O}(n)$. For the lower bound, note that our MCMC model implies that the variance of Y_i is larger than V because of the uncertainty of θ_i. Actually, under the MCMC model, conditional on the parameter A, the variance of the data $\{Y_i\}$ equals $V + A$. Therefore, the assumption in Eq. (2.14) is just to assume the observed data is not abnormal under the MCMC model when n is large enough. Note that only the existence of δ is required for establishing our main results. More precisely, the existence of δ is needed to obtain an upper bound for $\pi(R_0^c)$. If such δ does not exist, the MCMC model is (seriously) misspecified so the posterior distribution of the parameter A, which corresponds to the variance of a Normal distribution, may concentrate on 0. In that case, our upper bound on $\pi(R_0^c)$ does not hold.

Then we show that, under the assumption Eq. (2.14), with initial state

$$\bar{\theta}^{(0)} = \bar{Y}, \quad A^{(0)} = \begin{cases} \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} - V, & \text{if } \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} > V, \\ \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}, & \text{otherwise,} \end{cases} \tag{2.15}$$

and $\mu^{(0)}$ arbitrary (since $\mu^{(0)}$ will be updated in the first step of the Gibbs sampler), the mixing time of the Gibbs sampler to guarantee small total variation distance to stationarity is bounded by some constant when n is large enough.

2.2.1 Main Results

First, we obtain a quantitative bound for large enough n, which is given in the following theorem.

Theorem 2.2.2. *Under the assumption Eq. (2.14), with initial state Eq. (2.15), there exists a positive integer N which does not depend on k, some constants $C_1 > 0, C_2 > 0, C_3 > 0$ and $0 < \gamma < 1$, such that for all $n \geq N$ and for all k, we have*

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq C_1 \gamma^k + C_2 \frac{k(1+k)}{n} + C_3 \frac{k}{\sqrt{n}}. \tag{2.16}$$

Proof. We first choose the drift function, which is given in the following lemma.

Lemma 2.2.3. *Let $\Delta = \sum_{i=1}^n (Y_i - \bar{Y})^2$ and $x = (A, \mu, \theta_1, \ldots, \theta_n)$. Define the drift function $f_n(x)$ by*

$$f_n(x) := n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2. \tag{2.17}$$

Let $x^{(k)} = (A^{(k)}, \mu^{(k)}, \theta_1^{(k)}, \ldots, \theta_n^{(k)})$ be the state of the Markov chain at the k-th iteration, then we have

$$\mathbb{E}[f_n(x^{(k+1)}) \,|\, x^{(k)}] \leq \left(\frac{V^2 + 2VA^{(k)}}{V^2 + 2VA^{(k)} + (A^{(k)})^2}\right)^2 f_n(x^{(k)}) + b, \quad \forall x^{(k)} \in \mathcal{X} \tag{2.18}$$

where $b = \mathcal{O}(1)$.

Proof. See Section 2.3.3. □

Note that in Eq. (2.18), the term $\left(\frac{V^2+2VA^{(k)}}{V^2+2VA^{(k)}+(A^{(k)})^2}\right)^2$ depends on the coordinate $A^{(k)}$ of the state $x^{(k)}$ and is not bounded away from 1, since $A^{(k)}$ can be arbitrarily close to 0. Therefore, $\left(\frac{V^2+2VA^{(k)}}{V^2+2VA^{(k)}+(A^{(k)})^2}\right)^2$ cannot be bounded by some λ such that $0 < \lambda < 1$ and we cannot directly establish the traditional drift condition Eq. (2.2) by Eq. (2.18). In the following, we establish the generalized drift condition Definition 2.1.1 using a "large set".

According to Eq. (2.14), for large enough n, we have $\frac{\Delta}{n-1} > V$. Then, we choose a threshold T such that, for large enough n, we have $0 < T < \frac{\Delta}{n-1} - V$. Defining $\lambda_T := \left(\frac{V^2+2VT}{V^2+2VT+T^2}\right)^2 < 1$, we get

$$\mathbb{E}[f_n(x^{(k+1)}) \,|\, x^{(k)}] \leq \lambda_T \, f_n(x^{(k)}) + b, \quad \forall x \in R_T. \tag{2.19}$$

where the "large set", R_T, is defined by

$$R_T := \left\{ x \in \mathcal{X} : \left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2 \leq \left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2 \right\}. \tag{2.20}$$

In order to satisfy the new drift condition in Definition 2.1.1, we verify (C1'). Note that in our example the transition kernel of the Gibbs sampler can be written as a composition of reversible steps and only the last step of the Gibbs sampler updates the parameter A which is used for defining the "large set" R_T. Therefore, in order to verify Eq. (2.6), it suffices to check the last step if the value of the drift function increases by updating $x^{(k)} \in R_T$ to $x^{(k+1)} \in R_T^c$. By the definition of R_T, we have

$$\begin{aligned} \left[\left(\frac{\Delta}{n-1} - V\right) - A^{(k)}\right]^2 &\leq \left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2, \quad \forall x^{(k)} \in R_T \\ \left[\left(\frac{\Delta}{n-1} - V\right) - A^{(k+1)}\right]^2 &> \left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2, \quad \forall x^{(k+1)} \notin R_T. \end{aligned} \tag{2.21}$$

This implies the value of $f_n(x)$ increases if the Markov chain is outside of the "large set" after updating A. Therefore, the generalized drift condition in Definition 2.1.1 is satisfied.

Now we can use Theorem 2.1.4 to derive a quantitative bound for the Gibbs sampler. We first present some useful lemmas.

Lemma 2.2.4. *If $T = \Theta(1)$, by choosing the size of the "small set" $R = \{x \in \mathcal{X} : f_n(x) \leq d\}$ to satisfy $d = \mathcal{O}(1)$ and $d > \frac{b}{1-\lambda_T}$, there exists a probability measure $Q(\cdot)$ such that the Markov chain satisfies a minorization condition in Eq. (2.7) with the minorization volumne $\epsilon = \Theta(1)$.*

Proof. See Section 2.3.4. □

Lemma 2.2.5. *With the initial state given by Eq. (2.15), there exists a positive integer N, which does*

not depend on k, such that for all $n \geq N$, we have

$$\begin{aligned} & k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \\ & \leq \frac{k}{\sqrt{n}} \frac{\sqrt{b}(2V/\delta + 1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|} + \frac{k(1+k)}{2n} \frac{b}{\left[\left(\frac{\Delta}{n-1} - V\right) - T\right]^2}. \end{aligned} \tag{2.22}$$

Proof. See Section 2.3.5. □

Now we derive a quantitative bound for the Gibbs sampler for large enough n by combing results together. First, from Lemma 2.2.3, we have $b = \mathcal{O}(1)$. Recall that $\lambda_T = \left(\frac{V^2+2VT}{V^2+2VT+T^2}\right)^2$. We obtain $\frac{b}{1-\lambda_T} = \mathcal{O}(1)$ by choosing $T = \Theta(1)$. Since $d > \frac{b}{1-\lambda_T}$, we can choose the size of small set to be $d = \mathcal{O}(1)$. Then by Lemma 2.2.4, we obtain the minorization volume $\epsilon = \Theta(1)$. For $Q(R_T)$, we know from Lemma 2.2.5 that $P(x^{(0)}, R_T^c) = \mathcal{O}(1/n)$, where $x^{(0)} \in R$. This implies that $\epsilon Q(R_T^c) = \mathcal{O}(1/n)$. Since $\epsilon = \Theta(1)$, we have $\epsilon Q(R_T) = \epsilon - \epsilon Q(R_T^c) = \Theta(1)$. Furthermore, by definition $\alpha^{-1} = \frac{1+2b+\lambda_T d}{1+d} < 1$, it can be verified that α^{-1} is bounded away from 0 when $T = \Theta(1)$ and $d = \mathcal{O}(1)$. Next, since $\Lambda = 1+2(\lambda_T d+b) = \Theta(1)$, ignoring the term α^{rk} in Eq. (2.8), we choose $r = \log(\alpha)/\log(\alpha\Lambda/(1-\epsilon Q(R_T)))$ to balance the order of $(1-\epsilon Q(R_T))^r$ and $\alpha^{-1}(\alpha\Lambda)^r$ and define $\gamma := (1-\epsilon Q(R_T))^r = \alpha^{-1}(\alpha\Lambda)^r$. Then we have $\gamma = \Theta(1)$ and $0 < \gamma < 1$. Furthermore, since $f_n(x^{(0)}) = 0$ for large enough n and $\frac{b}{1-\lambda_T} = \mathcal{O}(1)$, we can pick a constant C_1 such that $C_1 \geq 2 + \frac{b}{1-\lambda_T}$ for large enough n. Finally, we have $k\pi(R_T^c) + \sum_{i=1}^k P^i(x^{(0)}, R_T^c) \leq C_2 \frac{k(1+k)}{n} + C_3 \frac{k}{\sqrt{n}}$ by Lemma 2.2.5, then Theorem 2.2.2 follows from Theorem 2.1.4. □

Next, we translate the quantitative bound in Theorem 2.2.2 into the convergence complexity in terms of mixing time. We show the convergence complexity is $\mathcal{O}(1)$. Intuitively, to make the term $C_1\gamma^k$ in Eq. (2.16) arbitrarily small, k needs to have a complexity order of $\mathcal{O}(1)$ since γ does not depend on n. The residual terms $C_2\frac{k(1+k)}{n} + C_3\frac{k}{\sqrt{n}} \to 0$ when $k = o(\sqrt{n})$. Therefore, the complexity bound on the mixing time of the Gibbs sampler equals the smaller complexity order between $\mathcal{O}(1)$ and $o(\sqrt{n})$, which is $\mathcal{O}(1)$. The formal result is given in the following.

Theorem 2.2.6. *For any $0 < c < 1$, recall the definition of the mixing time K_c in Definition 2.1.10. We write K_c as $K_c(n)$ to emphasize its dependence on n. Under the assumptions of Theorem 2.2.2, there exists $N_c = \Theta(1)$ and $\bar{K}_c = \Theta(1)$ such that*

$$K_c(n) \leq \bar{K}_c, \quad \forall n \geq N_c. \tag{2.23}$$

Proof. See Section 2.3.2. □

2.2.2 Discussions

We give further comments and discussions on the analysis of the Gibbs sampler.

Drift function

In the proof of Theorem 2.2.2, we have used a "centered" drift function shown in Eq. (2.17). To check this, we select a "typical" state $\tilde{x} = (\tilde{A}, \tilde{\mu}, \tilde{\theta}_1, \ldots, \tilde{\theta}_n)$ such that $\tilde{\theta}_i = Y_i$ and $\tilde{A} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$ then the

scaled drift function $f_n(\tilde{x})/n = nV^2/n = \Theta(1)$. We then hope to establish b such that $b/n = o(1)$, or equivalently, $b = o(n)$. Indeed, the established generalized drift condition has $b = \mathcal{O}(1) = o(n)$, which implies the drift function is "centered".

"Large set"

The result in Eq. (2.18) implies that those states whose value of A are close to zero are "bad" states. Therefore, the goal of choosing the "larget set" in Eq. (2.20) is to ruling out those states. Note that we have applied the trick that ruling more states with "high energy" could make Eq. (2.6) easier to establish. In the "large set" R_T defined by Eq. (2.20), we have also ruled out the states x whose value of A are larger than $\left|\left(\frac{\Delta}{n-1} - V\right) - T\right| + \left(\frac{\Delta}{n-1} - V\right)$. Note that these states are not "bad" states. However, by ruling out them, it is easy to establish Eq. (2.6) as shown in the proof of Theorem 2.2.2.

The upper bound in Eq. (2.22)

Although the upper bound of $k\,\pi(R_T^c) + \sum_{i=1}^k P^i(x^{(0)}, R_T^c)$ shown in Eq. (2.22) is loose, it is already enough for showing the mixing time of the Gibbs sampler is $\mathcal{O}(1)$. The proof of Lemma 2.2.5 only makes use of the form of drift function and the definition of "large set", and does not depend on the particular form of the transition kernel of the Gibbs sampler. We expect that, in general, tighter upper bound on $k\,\pi(R_T^c) + \sum_{i=1}^k P^i(x^{(0)}, R_T^c)$ could be obtained, depending on the choice of "large set" and the MCMC algorithm to be analyzed. This may involve carefully bounding the tail probability of the transition kernel.

Initial state

The main results in Theorem 2.2.2 and Theorem 2.2.6 hold for a particular initial state given in Eq. (2.15). We discuss other initial states than the one given in Eq. (2.15). Note that the new bound in Lemma 2.2.3 holds for any initial state that is in the "large set". Therefore, we can extend the results in Theorem 2.2.2 to get bounds when the Markov chain starts from some other initial states in the "large set". Recall the assumption on the observed data $\{Y_i\}$ in Eq. (2.14), we have assumed there exists $\delta > 0$ such that $\frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} \geq V + \delta$ for large enough n. Note that the existence of such δ is sufficient to obtain the results in Theorem 2.2.2 and Theorem 2.2.6. In order to get bounds when the MCMC algorithm starts from other initial states, we assume δ is known and establish upper bounds using δ explicitly. We define the "large set" Eq. (2.20) using $T = \delta$ and the extension of Theorem 2.2.2 is given in the following.

Theorem 2.2.7. *Let $\Delta = \sum_{i=1}^n (Y_i - \bar{Y})^2$. Under the assumption Eq. (2.14), if the Markov chain starts with any initial state $x^{(0)} \in R_\delta$ (defined in Eq. (2.20) with $T = \delta$), there exists a positive integer N, which does not depend on k, some constants $C_1 > 0, C_2 > 0, C_3 > 0, C_4 > 0$ and $0 < \gamma < 1$, such that for all $n \geq N$ and for all k, we have*

$$\|\mathcal{L}(X^{(k)}) - \pi\|_{\text{var}} \leq [C_1 + f_n(x^{(0)})]\gamma^k + C_2 \frac{k(1+k)}{n} + C_3 \frac{k}{\sqrt{n}} + C_4 f_n(x^{(0)}) \frac{k}{n}, \tag{2.24}$$

where $f_n(\cdot)$ is the drift function defined in Eq. (2.17).

Proof. Following the same proof of Theorem 2.2.2 by keeping the term $f_n(x^{(0)})$, the first two terms of the upper bound given in Eq. (2.8) can be replaced by $[C_1 + f_n(x^{(0)})]\gamma^k$ and the last term of the upper bound in Eq. (2.8) can be replaced by $\sum_{i=1}^k P^i(x^{(0)}, R_\delta^c) \leq C_2 \frac{k(1+k)}{n} + C_4 f_n(x^{(0)}) \frac{k}{n}$. □

From Theorem 2.2.7, we can immediately obtain a complexity bound when the Markov chain starts within a subset of the "large set", which is given in the following. This result suggests that if the Markov chain starts from an initial state which is not "too far" from the state given in Eq. (2.15), the Markov chain still mixes fast. The mixing time becomes $\mathcal{O}(\log n)$ instead of $\mathcal{O}(1)$.

Corollary 2.2.8. *Under the assumption Eq. (2.14), if the initial state of the Markov chain satisfies* $x^{(0)} \in \{x \in R_\delta : f_n(x) = o(n/\log n)\}$, *the mixing time of the Gibbs sampler is* $\mathcal{O}(\log n)$.

Note that $\{x \in R_\delta : f_n(x) = o(n/\log n)\}$ defines a subset of the "large set" R_δ, and the above result shows that the mixing time is $\mathcal{O}(\log n)$ if the initial state is in this subset. The order $o(n/\log n)$ comes from a balance between $f_n(x^{(0)})\gamma^k$ and $f_n(x^{(0)})\frac{k}{n}$. We conjecture the same complexity order of $\mathcal{O}(\log n)$ on the mixing time may hold even if the initial state is in a larger subset, for example $\left\{x^{(0)} \in R_\delta : f_n(x^{(0)}) = \Theta(n)\right\}$. However, in order to prove this, we need to derive tighter upper bound of $\sum_{i=1}^{k} P^i(x^{(0)}, R_\delta^c)$ which is a non-trivial task. We therefore leave it as an open problem.

Finally, we do not have upper bounds for the Markov chain when the initial state is outside of the "large set" since the new bound in Theorem 2.1.4 requires the Markov chain starts within the "large set". For this particular Gibbs sampler example, numerical experiments suggest that, if the Markov chain starts from a "bad" state, the number of iterations required for the Markov chain to mix can be much larger than $\mathcal{O}(\log n)$. In high-dimensional settings, when the dimension of the state space goes to infinity, the Markov chain may not mix fast starting from any state. This observation is loosely consistent with various observations made by Hairer, Mattingly, and Scheutzow [HMS11].

The constants in Theorem 2.2.2

In Theorem 2.2.2, we do not compute the constants N, C_1, C_2, and C_3 explicitly. Actually, C_2 is given explicitly in Lemma 2.2.5. C_3 is given in Lemma 2.2.5 but it depends on the unknown constant $\delta > 0$ from the assumption Eq. (2.14). Furthermore, C_1 can be explicitly computed under much more tedious computations. Finally, N depends on the unknown constant N_0 in Eq. (2.14) and the resulting concentration property of the posterior distribution for parameter A by Eq. (2.14). Therefore, if we make stronger assumptions on the observed data $\{Y_i\}$, it is then possible to compute all the constants in Theorem 2.2.2 explicitly under tedious computations, though we do not pursue that here.

Relation to spectral gaps

Many approaches in MCMC literature bound the spectral gap of the corresponding Markov operator[LV03; Vem05; LV06; WSH09a; WSH09b]. However, on general state spaces, the spectral gap is zero for Markov chains which are not geometrically ergodic, even if they do converge to stationarity. Our results do not require the Markov chain to be geometrically ergodic. Instead, we only require the constructed "restricted" chain on the "large set" in our proof is geometrically ergodic. Therefore, we cannot connect our results to bounds on spectral gaps. Furthermore, we do not require the Markov chain to be reversible. So our results apply even in the non-reversible cases, which makes spectral gaps harder to study or interpret. For these reasons, we do not present the main results in terms of spectral gaps.

2.3 Appendices

2.3.1 Proof of Theorem 2.1.4

Recall that R denotes the "small set" and R_0 denotes the "large set". We first construct a transition kernel for a "restricted" chain define on R_0, $\tilde{P}(x,\cdot), \forall x \in R_0$. One goal of this construction is that the stationary distribution of the kernel $\tilde{P}$ equals to the $\pi(\cdot)$ restricted on the "large set" R_0, i.e., $\pi'(\mathrm{d}x) := \pi(\mathrm{d}x)/\pi(R_0), \forall x \in R_0$. We consider two different constructions depending on (C1) or (C1') in Definition 2.1.1 holds.

- If (C1) in Definition 2.1.1 holds, then we define the kernel $\tilde{P}$ as the transition kernel of the "trace chain" constructed as follows. Let $X^{(m)}$ be a Markov chain with kernel P, we define a sequence of random entrance time $\{m_i\}_{i\in\mathbb{N}}$ by $m_0 := \min\{m \geq 0 : X^{(m)} \in R_0\}$, $m_i := \min\{m > m_{i-1} : X^{(m)} \in R_0\}$. Then $\{X^{(m_i)}\}_{i\in\mathbb{N}}$ is the "trace chain" and the transition kernel $\tilde{P}(x,B) := \mathbb{P}(X^{(m_1)} \in B \,|\, X^{(m_0)} = x), \forall x \in R_0$. It is clear that the "trace chain" is obtained by "stopping the clock" when the original chain is outside R_0, the constructed $\tilde{P}$ is a valid transition kernel. It can be verified that the stationary distribution of this "trace chain" is π'.

- If (C1') in Definition 2.1.1 holds, then we construct the "restricted chain" using the kernel $\tilde{P} = \prod_{i=1}^{I} \tilde{P}_i$ where $\tilde{P}_i(x,\mathrm{d}y) := P_i(x,\mathrm{d}y)$ for $x, y \in R_0, x \neq y$, and $\tilde{P}_i(x,x) := 1 - P_i(x, R_0\backslash\{x\}), \forall x \in R_0$. Note that since each P_i is reversible, one can easily verify that each $\tilde{P}_i$ is also reversible and the stationary distribution of $\tilde{P}$ is π'.

Suppose that $X^{(m)}$ and $Y^{(m)}$ are two realizations of the Markov chain, where $X^{(m)}$ starts with the initial distribution $\nu(\cdot)$ and $Y^{(m)}$ starts with the stationary distribution $\pi(\cdot)$. We define $\tilde{X}^{(m)}$ and $\tilde{Y}^{(m)}$ to be two realizations of a constructed "restricted" Markov chain on the "large set" with the transition kernel $\tilde{P}(x,\cdot), \forall x \in R_0$. We assume $\tilde{X}^{(m)}$ starts with the same initial distribution $\nu(\cdot)$ as $X^{(m)}$ and $\tilde{Y}^{(m)}$ starts with $\pi'(\cdot)$. Since $\nu(R_0) = 1$, we assume $X^{(0)} = \tilde{X}^{(0)}$. This rest of the proof is a modification of the original proof of the drift-and-minorization method using coupling in [Ros95a].

We define the hitting times of $(\tilde{X}^{(m)}, \tilde{Y}^{(m)})$ to $R \times R$ as follows.

$$\begin{aligned} t_1 &:= \inf\{m \geq 0 : (\tilde{X}^{(m)}, \tilde{Y}^{(m)}) \in R \times R\}, \\ t_i &:= \inf\{m \geq t_{i-1} + 1 : (\tilde{X}^{(m)}, \tilde{Y}^{(m)}) \in R \times R\}, \quad \forall i > 1. \end{aligned} \tag{2.25}$$

Let $N_k := \max\{i : t_i < k\}$. Then N_k denotes the number of $(\tilde{X}^{(m)}, \tilde{Y}^{(m)})$ to hit $R \times R$ in the first k iterations. The following result gives an upper bound for $\|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\mathrm{var}}$.

Lemma 2.3.1. *When the Markov chain satisfies the minorization condition in Eq. (2.7), for any $j > 0$, we have*

$$\begin{aligned} \|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\mathrm{var}} \leq & (1 - \epsilon Q(R_0))^j + \mathbb{P}(N_k < j) \\ & + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c). \end{aligned} \tag{2.26}$$

Proof. First, by triangle inequality

$$\begin{aligned}\|\mathcal{L}(X^{(k)}) - \mathcal{L}(Y^{(k)})\|_{\text{var}} \leq \|\mathcal{L}(\tilde{X}^{(k)}) - \mathcal{L}(\tilde{Y}^{(k)})\|_{\text{var}} + \|\mathcal{L}(X^{(k)}) - \mathcal{L}(\tilde{X}^{(k)})\|_{\text{var}} \\ + \|\mathcal{L}(Y^{(k)}) - \mathcal{L}(\tilde{Y}^{(k)})\|_{\text{var}}.\end{aligned} \tag{2.27}$$

By the coupling inequality $\|\mathcal{L}(X^{(k)}) - \mathcal{L}(\tilde{X}^{(k)})\|_{\text{var}} \leq \mathbb{P}(X^{(k)} \neq \tilde{X}^{(k)}) \leq \sum_{m=1}^{k} \mathbb{P}(X^{(m)} \notin R_0)$, we have

$$\begin{aligned}&\|\mathcal{L}(Y^{(k)}) - \mathcal{L}(\tilde{Y}^{(k)})\|_{\text{var}} + \|\mathcal{L}(X^{(k)}) - \mathcal{L}(\tilde{X}^{(k)})\|_{\text{var}} \\ &\leq \sum_{m=1}^{k} \mathbb{P}\left(Y^{(m)} \notin R_0\right) + \sum_{m=1}^{k} \mathbb{P}\left(X^{(m)} \notin R_0\right) \\ &\leq k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c).\end{aligned} \tag{2.28}$$

Finally, the Markov chain with kernel $\tilde{P}(x, \cdot)$ satisfies both drift condition

$$\mathbb{E}(f(\tilde{X}^{(1)}) \,|\, \tilde{X}^{(0)} = x) \leq \lambda f(x) + b, \quad \forall x \in R_0, \tag{2.29}$$

and minorization condition

$$\tilde{P}(x, \mathrm{d}y) \geq [\epsilon Q(R_0)] \frac{Q(\mathrm{d}y)}{Q(R_0)}, \quad \forall x, y \in R_0. \tag{2.30}$$

Using the result from [Ros95a, Theorem 1], we have

$$\|\mathcal{L}(\tilde{X}^{(k)}) - \mathcal{L}(\tilde{Y}^{(k)})\|_{\text{var}} \leq (1 - \epsilon Q(R_0))^j + \mathbb{P}(N_k < j). \tag{2.31}$$

□

Next, we further upper bound the term $\mathbb{P}(N_k < j)$ slightly tighter than [Ros95a]. Define the i-th gap of return times by $r_i := t_i - t_{i-1}, \forall i > 1$, then

Lemma 2.3.2. *For any $\alpha > 1$ and $j > 0$, and $k > j$,*

$$\mathbb{P}(N_k < j) \leq \frac{1}{\alpha^k - \alpha^j} \left[\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i} \right) - \alpha^j \right]. \tag{2.32}$$

Proof. Note that $\{N_k < j\} = \{t_j \geq k\} = \{r_1 + \cdots + r_j \geq k\}$ and $r_1 + \cdots + r_j \geq j$ by definition. Then the result comes from Markov's inequality

$$\begin{aligned}\mathbb{P}(N_k < j) &= \mathbb{P}(r_1 + \cdots + r_j \geq k) \\ &= \mathbb{P}(\alpha^{r_1 + \cdots + r_j} - \alpha^j \geq \alpha^k - \alpha^j) \\ &\leq \frac{1}{\alpha^k - \alpha^j} \left[\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i} \right) - \alpha^j \right].\end{aligned} \tag{2.33}$$

□

Next, we bound $\mathbb{E}\left(\prod_{i=1}^{j} \alpha^{r_i}\right)$ following the exact same arguments as in [Ros95a, Proof of Lemma 4

and Theorem 12], which gives

$$\mathbb{E}\left(\prod_{i=1}^{j}\alpha^{r_i}\right) \leq (\alpha\Lambda)^{j-1}\left[1+\mathbb{E}_{\nu}(f(x))+\mathbb{E}_{\pi'}(f(x))\right]. \tag{2.34}$$

By the drift condition for $\tilde{P}(x,\cdot)$ in Eq. (2.29), taking expectations on both sides of Eq. (2.29) leads to $\mathbb{E}_{\pi'}(f(x)) \leq \frac{b}{1-\lambda}$. Therefore, setting $j = rk+1$ and combining all results together yields

$$\begin{aligned}\|\mathcal{L}(X^{(k)})-\pi\|_{\text{var}} \leq (1-\epsilon Q(R_0))^{rk+1} + \frac{(\alpha\Lambda)^{rk}\left[1+\mathbb{E}_{\nu}(f(x))+\frac{b}{1-\lambda}\right]-\alpha^{rk+1}}{\alpha^{k}-\alpha^{rk+1}} \\ + k\,\pi(R_0^c) + \sum_{i=1}^{k} P^i(\nu, R_0^c).\end{aligned} \tag{2.35}$$

Finally, we slightly relax the upper bound by replacing α^{rk+1} with α^{rk} in both the denominator and numerator. Then Theorem 2.1.4 is proved by further relaxing $(1-\epsilon)^{rk+1}$ to $(1-\epsilon)^{rk}$.

2.3.2 Proof of Theorem 2.2.6

Using Theorem 2.2.2, one sufficient condition for

$$\|\mathcal{L}(X^{(k)})-\pi\|_{\text{var}} \leq c \tag{2.36}$$

is that $n \geq N$ and

$$C_1\gamma^k \leq \frac{c}{3}, \quad C_2\frac{(1+k)^2}{n} \leq \frac{c}{3}, \quad C_3\frac{k}{\sqrt{n}} \leq \frac{c}{3}. \tag{2.37}$$

This requires the number of iterations, k, satisfies

$$\frac{\log(C_1)-\log(c/3)}{\log(1/\gamma)} \leq k \leq \min\left\{\sqrt{\frac{c/3}{C_3}}\sqrt{n}-1, \frac{c/3}{C_3}\sqrt{n}\right\}. \tag{2.38}$$

Note that any k (if exists) satisfying the above equation provides an upper bound for the mixing time $K_c(n)$.

That is, for any $n \geq N$ such that

$$\frac{\log(C_1)-\log(c/3)}{\log(1/\gamma)} \leq \min\left\{\sqrt{\frac{c/3}{C_3}}\sqrt{n}-1, \frac{c/3}{C_3}\sqrt{n}\right\}, \tag{2.39}$$

which is equivalent to

$$n \geq \max\left\{N, \left[\bar{K}_c\frac{3C_3}{c}\right]^2, \left[\left(\bar{K}_c+1\right)\sqrt{\frac{3C_3}{c}}\right]^2\right\} =: N_c, \tag{2.40}$$

we have $\bar{K}_c := \frac{\log(C_1)-\log(c)+\log(3)}{\log(1/\gamma)}$ is an upper bound of the mixing time.

Finally, it can be seen that both $\bar{K}_c = \Theta(1)$ and $N_c = \Theta(1)$.

2.3.3 Proof of Lemma 2.2.3

In this proof, we write $f_n(x)$ as $f(x)$ for simplicity. Recall that the order of Gibbs sampling for computing the first scan is:

$$\begin{aligned}
\mu^{(1)} &\sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right), \\
\theta_i^{(1)} &\sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right), \\
A^{(1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right).
\end{aligned} \tag{2.41}$$

It suffices to show that for $\Delta = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ and

$$f(x) = n(\bar{\theta} - \bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2, \tag{2.42}$$

we have

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] \leq \left(\frac{V^2 + 2VA^{(0)}}{V^2 + 2VA^{(0)} + (A^{(0)})^2}\right)^2 f(x^{(0)}) + b, \tag{2.43}$$

where $b = \mathcal{O}(1)$.

Note that we can compute the expectation in $\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}]$ by three steps, according to the reverse order of the Gibbs sampling. To simplify the notation, we define σ-algebras that we condition on:

$$\begin{aligned}
\mathcal{G}_A &:= \sigma(A^{(0)}, \{\theta_i^{(1)}\}, \mu^{(1)}), \\
\mathcal{G}_\theta &:= \sigma(A^{(0)}, \{\theta_i^{(0)}\}, \mu^{(1)}), \\
\mathcal{G}_\mu &:= \sigma(A^{(0)}, \{\theta_i^{(0)}\}, \mu^{(0)}).
\end{aligned} \tag{2.44}$$

Then we have

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] = \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_\mu] = \mathbb{E}[\mathbb{E}[\mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \,|\, \mathcal{G}_\theta] \,|\, \mathcal{G}_\mu]. \tag{2.45}$$

The three steps are as follows:

1. Compute the expectation over $A^{(1)}$ given $\{\theta_i^{(1)}\}$ and $\mu^{(1)}$. This is to compute the conditional expectation

$$f'(x^{(1)}) := \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A], \tag{2.46}$$

where we write $\mathbb{E}[\cdot \,|\, \mathcal{G}_A]$ to denote the the expectation is over (recall that a and b are constants from the prior $\mathbf{IG}(a, b)$)

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right) \tag{2.47}$$

for given $\theta^{(1)}$ and $\mu^{(1)}$.

2. Compute the expectation over $\{\theta_i^{(1)}\}$ given $\mu^{(1)}$. This is to compute the conditional expectation

$$f''(x^{(1)}) := \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta], \tag{2.48}$$

where we use $\mathbb{E}[\cdot \,|\, \mathcal{G}_\theta]$ to denote the expectation is over

$$\theta_i^{(1)} \sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right), \quad i = 1, \ldots, n, \tag{2.49}$$

for given $\mu^{(1)}$ and $A^{(0)}$.

3. Compute the expectation over $\mu^{(1)}$. This is to compute the conditional expectation

$$\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] = \mathbb{E}[f''(x^{(1)}) \,|\, \mathcal{G}_\mu], \tag{2.50}$$

where we have used $\mathbb{E}[\cdot \,|\, \mathcal{G}_\mu]$ to denote the expectation is over

$$\mu^{(1)} \sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right) \tag{2.51}$$

for given $\{\theta_i^{(0)}\}$ and $A^{(0)}$.

In the following, we compute the three steps, respectively. We use $\mathcal{O}(1)$ to denote terms that can be upper bounded by some constant that does not depend on the state.

Compute $f'(x^{(1)}) = \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A]$

The first term of $f(x^{(1)})$ is $n(\bar{\theta}^{(1)} - \bar{Y})^2$, which is $\mathcal{G}_A$-measurable by construction. Thus, $\mathbb{E}[n(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_A] = n(\bar{\theta}^{(1)} - \bar{Y})^2$. Then

$$\begin{aligned} f'(x^{(1)}) &= \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \\ &= n(\bar{\theta}^{(1)} - \bar{Y})^2 + n\mathbb{E}\left\{\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(1)}\right]^2 \,|\, \mathcal{G}_A\right\}. \end{aligned} \tag{2.52}$$

Note that

$$\begin{aligned} &n\mathbb{E}\left\{\left[\left(\frac{\Delta}{n-1} - V\right) - A^{(1)}\right]^2 \,|\, \mathcal{G}_A\right\} \\ &= n\left(\frac{\Delta}{n-1} - V\right)^2 + n\mathbb{E}[(A^{(1)})^2 \,|\, \mathcal{G}_A] - 2n\left(\frac{\Delta}{n-1} - V\right)\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A]. \end{aligned} \tag{2.53}$$

Recall that $\mathbb{E}[\cdot \,|\, \mathcal{G}_A]$ denotes that the expectation is over

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n}(\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right), \tag{2.54}$$

where a and b are constants from the prior $\mathbf{IG}(a, b)$. The mean and variance of $A^{(1)}$ can be written in closed forms since $A^{(1)}$ follows from an inverse Gamma distribution. Denoting $S := \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$, we

can write the mean of $A^{(1)}$ using S as follows:

$$\begin{aligned}\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A] &= \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2 + 2b}{n - 1 + 2(a-1)} \\ &= \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1} + \frac{2b}{n-1+2(a-1)} \\ &\quad - \left(\frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1} \right) \left(\frac{2(a-1)}{n-1+2(a-1)} \right) \\ &= S + \mathcal{O}(1/n) + \mathcal{O}(1/n) S.\end{aligned} \tag{2.55}$$

Similarly, the variance of $A^{(1)}$ can be written in terms of S as well:

$$\begin{aligned}\mathrm{var}[A^{(1)} \,|\, \mathcal{G}_A] &= \frac{(\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2/2 + b)^2}{[(n-1)/2 + (a-1)][(n-1)/2 + (a-2)]} \\ &= \frac{1}{(n-1)/2 + (a-2)} \left(\mathbb{E}[A^{(1)} \,|\, \mathcal{G}_A] \right)^2 \\ &= \mathcal{O}(1/n) \left(S + \mathcal{O}(1/n) + \mathcal{O}(1/n) S \right)^2 \\ &= \mathcal{O}(1/n) S^2 + \mathcal{O}(1/n^2) S + \mathcal{O}(1/n^3).\end{aligned} \tag{2.56}$$

Substituting the mean and variance of $A^{(1)}$ in terms of S, we have

$$\begin{aligned}f'(x^{(1)}) &= \mathbb{E}[f(x^{(1)}) \,|\, \mathcal{G}_A] \\ &= n(\bar{\theta}^{(1)} - \bar{Y})^2 + n \left(\frac{\Delta}{n-1} - V \right)^2 + nS^2 - 2n \left(\frac{\Delta}{n-1} - V \right) S \\ &\quad + \mathcal{O}(1) + \mathcal{O}(1) S + \mathcal{O}(1) S^2.\end{aligned} \tag{2.57}$$

Compute $f''(x^{(1)}) = \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta]$

Note that the terms in $f'(x^{(1)})$ involving $\{\theta_i^{(1)}\}$ are $(\bar{\theta}^{(1)} - \bar{Y})^2$ and $S = \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$. Then

$$\begin{aligned}f''(x^{(1)}) &= \mathbb{E}[f'(x^{(1)}) \,|\, \mathcal{G}_\theta] \\ &= n\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta \right] + n \left(\frac{\Delta}{n-1} - V \right)^2 \\ &\quad + n\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] - 2n \left(\frac{\Delta}{n-1} - V \right) \mathbb{E}[S \,|\, \mathcal{G}_\theta] \\ &\quad + \mathcal{O}(1) + \mathcal{O}(1)\mathbb{E}[S \,|\, \mathcal{G}_\theta] + \mathcal{O}(1)\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta].\end{aligned} \tag{2.58}$$

Therefore, it suffices to compute the following terms

$$\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta \right], \quad \mathbb{E}[S \,|\, \mathcal{G}_\theta], \quad \mathbb{E}[S^2 \,|\, \mathcal{G}_\theta]. \tag{2.59}$$

Note that $\{\theta_i^{(1)}\}$ are independent (but not identically distributed) conditional on $\mathcal{G}_\theta$. For the first

term $\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right]$, we have

$$\begin{aligned}
\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right] &= \mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}} + \frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}} - \bar{Y}\right)^2 \,|\, \mathcal{G}_\theta\right] \\
&= \mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}}\right)^2 \,|\, \mathcal{G}_\theta\right] + \left(\frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}} - \bar{Y}\right)^2 \\
&\quad + 2\left(\frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}} - \bar{Y}\right)\mathbb{E}\left[\left(\bar{\theta}^{(1)} - \frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}}\right) \,|\, \mathcal{G}_\theta\right] \\
&= \mathrm{var}[\bar{\theta}^{(1)} \,|\, \mathcal{G}_\theta] + \left(\frac{V}{V+A^{(0)}}\right)^2\left(\mu^{(1)}-\bar{Y}\right)^2 \\
&= \frac{1}{n}\frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{V}{V+A^{(0)}}\right)^2\left(\mu^{(1)}-\bar{Y}\right)^2
\end{aligned} \tag{2.60}$$

For the other two terms involving S, we have the following lemma.

Lemma 2.3.3. *For $S = \frac{\sum_i(\theta_i^{(1)}-\bar{\theta}^{(1)})^2}{n-1}$, we have*

$$\mathbb{E}[S \,|\, \mathcal{G}_\theta] = \frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1}, \quad \mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathcal{O}(1/n). \tag{2.61}$$

Proof. Define $\eta_i := \theta_i^{(1)} - \frac{Y_iA^{(0)}}{V+A^{(0)}}$ then $\bar{\eta} = \bar{\theta}^{(1)} - \frac{\bar{Y}A^{(0)}}{V+A^{(0)}}$. Note that $\{\eta_i\}$ are i.i.d. conditional on $\mathcal{G}_\theta$ with

$$\eta_i \sim \mathcal{N}\left(\frac{\mu^{(1)}V}{V+A^{(0)}}, \frac{A^{(0)}V}{V+A^{(0)}}\right), \quad \bar{\eta} \sim \mathcal{N}\left(\frac{\mu^{(1)}V}{V+A^{(0)}}, \frac{1}{n}\frac{A^{(0)}V}{V+A^{(0)}}\right). \tag{2.62}$$

Next, we decompose $\sum_{i=1}^n(\theta_i^{(1)}-\bar{\theta}^{(1)})^2$ by

$$\begin{aligned}
\sum_{i=1}^n(\theta_i^{(1)}-\bar{\theta}^{(1)})^2 &= \sum_{i=1}^n\left(\eta_i + \frac{Y_iA^{(0)}}{V+A^{(0)}} - \bar{\eta} - \frac{\bar{Y}A^{(0)}}{V+A^{(0)}}\right)^2 \\
&= \sum_{i=1}^n\left((\eta_i-\bar{\eta})^2 + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2(Y_i-\bar{Y})^2 + \frac{2(\eta_i-\bar{\eta})(Y_i-\bar{Y})A^{(0)}}{V+A^{(0)}}\right).
\end{aligned} \tag{2.63}$$

Then we can obtain $\mathbb{E}[S \,|\, \mathcal{G}_\theta]$ by

$$\begin{aligned}
\mathbb{E}[S \,|\, \mathcal{G}_\theta] &= \mathbb{E}\left\{\left[\frac{\sum_i(\theta_i^{(1)}-\bar{\theta}^{(1)})^2}{n-1}\right] \,|\, \mathcal{G}_\theta\right\} \\
&= \mathbb{E}\left\{\left[\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1}\right] \,|\, \mathcal{G}_\theta\right\} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\sum_{i=1}^n(Y_i-\bar{Y})^2}{n-1} \\
&= \frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1}.
\end{aligned} \tag{2.64}$$

For $\mathrm{var}[S \,|\, \mathcal{G}_\theta]$, using the Cauchy-Schwartz inequality

$$\begin{aligned}
&\mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathbb{E}\left[(S - \mathbb{E}[S \,|\, \mathcal{G}_\theta])^2 \,|\, \mathcal{G}_\theta\right] \\
&= \mathbb{E}\left[\left(\frac{\sum_{i=1}^n(\eta_i - \bar{\eta})^2}{n-1} - \mathbb{E}_{\{\eta_i\}}\left[\frac{\sum_i(\eta_i - \bar{\eta})^2}{n-1}\right] + 2\frac{A^{(0)}}{V + A^{(0)}}\frac{\sum_{i=1}^n(\eta_i - \bar{\eta})(Y_i - \bar{Y})}{n-1}\right)^2 \,|\, \mathcal{G}_\theta\right] \\
&\leq 2\,\mathrm{var}\left[\frac{\sum_i(\eta_i - \bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right] + 8\left(\frac{A^{(0)}}{V + A^{(0)}}\right)^2 \frac{\mathbb{E}\left\{\left[\sum_i(\eta_i - \bar{\eta})(Y_i - \bar{Y})\right]^2 \,|\, \mathcal{G}_\theta\right\}}{(n-1)^2}.
\end{aligned} \tag{2.65}$$

Note that $\{\eta_i\}$ are i.i.d conditional on $\mathcal{G}_\theta$, we know

$$\mathbb{E}\left\{\left[\frac{\sum_i(\eta_i - \bar{\eta})^2}{n-1}\right]^2 \,|\, \mathcal{G}_\theta\right\} = \left\{\mathbb{E}\left[\frac{\sum_i(\eta_i - \bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right]\right\}^2 + \mathcal{O}(1/n). \tag{2.66}$$

That is, $\mathrm{var}\left[\frac{\sum_i(\eta_i - \bar{\eta})^2}{n-1} \,|\, \mathcal{G}_\theta\right] = \mathcal{O}(1/n)$. Finally, the term

$$\begin{aligned}
&\frac{\mathbb{E}\left\{\left[\sum_i(\eta_i - \bar{\eta})(Y_i - \bar{Y})\right]^2 \,|\, \mathcal{G}_\theta\right\}}{(n-1)^2} \\
&= \frac{\mathbb{E}\left\{\left[\sum_i(\eta_i - \bar{\eta})^2(Y_i - \bar{Y})^2\right] \,|\, \mathcal{G}_\theta\right\} + \mathbb{E}[\bar{\eta}^2 \,|\, \mathcal{G}_\theta]\sum_{i\neq j}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(n-1)^2} \\
&= \frac{\sum_i(Y_i - \bar{Y})^2}{(n-1)^2}\mathbb{E}\left[(\eta_1 - \bar{\eta})^2 \,|\, \mathcal{G}_\theta\right] + \mathcal{O}(1/n) \\
&= \frac{\Delta}{(n-1)^2}\frac{(n-1)\frac{A^{(0)}V}{V+A^{(0)}}}{n} + \mathcal{O}(1/n) = \mathcal{O}(1/n).
\end{aligned} \tag{2.67}$$

Therefore, we have $\mathrm{var}[S \,|\, \mathcal{G}_\theta] = \mathcal{O}(1/n)$. □

Next, using the following results

$$\begin{aligned}
\mathbb{E}[S \,|\, \mathcal{G}_\theta] &= \frac{A^{(0)}V}{V + A^{(0)}} + \left(\frac{A^{(0)}}{V + A^{(0)}}\right)^2 \frac{\Delta}{n-1} \\
&\leq V + \left(\frac{A^{(0)}}{V + A^{(0)}}\right)^2 \frac{\Delta}{n-1} = \mathcal{O}(1), \\
\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] &= \left(\mathbb{E}[S \,|\, \mathcal{G}_\theta]\right)^2 + \mathcal{O}(1/n) = \mathcal{O}(1),
\end{aligned} \tag{2.68}$$

we can first write $f''(x^{(1)})$ by

$$\begin{aligned}
f''(x^{(1)}) =& n\mathbb{E}\left[(\bar{\theta}^{(1)} - \bar{Y})^2 \,|\, \mathcal{G}_\theta\right] + n\left(\frac{\Delta}{n-1} - V\right)^2 \\
&+ n\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] - 2n\left(\frac{\Delta}{n-1} - V\right)\mathbb{E}[S \,|\, \mathcal{G}_\theta] + \mathcal{O}(1).
\end{aligned} \tag{2.69}$$

Then, using

$$\begin{aligned} n\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right] &= \frac{A^{(0)}V}{V+A^{(0)}} + n\left(\frac{V}{V+A^{(0)}}\right)^2 \left(\mu^{(1)}-\bar{Y}\right)^2 \\ &\leq V + \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} \end{aligned} \tag{2.70}$$

we further bound the terms

$$\begin{aligned} & n\mathbb{E}\left[(\bar{\theta}^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\theta\right] + n\left(\frac{\Delta}{n-1}-V\right)^2 \\ & \quad + n\mathbb{E}[S^2 \,|\, \mathcal{G}_\theta] - 2n\left(\frac{\Delta}{n-1}-V\right)\mathbb{E}[S \,|\, \mathcal{G}_\theta] \\ & \leq \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + n\left[\left(\frac{\Delta}{n-1}-V\right) - \mathbb{E}[S \,|\, \mathcal{G}_\theta]\right]^2 \\ & = \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + n\left[\frac{A^{(0)}V}{V+A^{(0)}} + \left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2\frac{\Delta}{n-1} - \left(\frac{\Delta}{n-1}-V\right)\right]^2 \\ & = \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + n\left[\frac{\Delta}{n-1}\left[\left(\frac{A^{(0)}}{V+A^{(0)}}\right)^2 - 1\right] + \left(\frac{A^{(0)}V}{V+A^{(0)}}+V\right)\right]^2 \\ & = \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + n\left(\frac{A^{(0)}}{V+A^{(0)}}+1\right)^2\left[\frac{\Delta}{n-1}\left(\frac{-V}{V+A^{(0)}}\right)+V\right]^2 \\ & = \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2. \end{aligned} \tag{2.71}$$

Finally, combing all the results yields

$$f''(x^{(1)}) = \frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2} + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2 + \mathcal{O}(1). \tag{2.72}$$

Compute $\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] = \mathbb{E}[f''(x^{(1)}) \,|\, \mathcal{G}_\mu]$

Recall that the expectation $\mathbb{E}[\cdot \,|\, \mathcal{G}_\mu]$ is over

$$\mu^{(1)} \sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right). \tag{2.73}$$

In the obtained expression of $f''(x^{(1)})$ from previous step, the only term involves $\mu^{(1)}$ is $\frac{nV^2\left(\mu^{(1)}-\bar{Y}\right)^2}{(V+A^{(0)})^2}$. Since

$$\mathbb{E}\left[(\mu^{(1)}-\bar{Y})^2 \,|\, \mathcal{G}_\mu\right] = (\bar{\theta}^{(0)}-\bar{Y})^2 + A^{(0)}/n, \tag{2.74}$$

we have

$$\begin{aligned}\mathbb{E}[f(x^{(1)}) \,|\, x^{(0)}] &= \mathbb{E}[f''(x^{(1)}) \,|\, \mathcal{G}_\mu] \\ &\leq \frac{nV^2}{(V+A^{(0)})^2}\left((\bar{\theta}^{(0)}-\bar{Y})^2 + \frac{A^{(0)}}{n}\right) \\ &\quad + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2 + \mathcal{O}(1) \\ &= \frac{nV^2(\bar{\theta}^{(0)}-\bar{Y})^2}{(V+A^{(0)})^2} \\ &\quad + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2 + \mathcal{O}(1).\end{aligned} \tag{2.75}$$

Finally, we complete the proof by

$$\begin{aligned}&\frac{nV^2(\bar{\theta}^{(0)}-\bar{Y})^2}{(V+A^{(0)})^2} + \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2 + \mathcal{O}(1) \\ &= \frac{nV^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left\{\frac{(V+A^{(0)})^2}{(V+2A^{(0)})^2}(\bar{\theta}^{(0)}-\bar{Y})^2 + \left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2\right\} + \mathcal{O}(1) \\ &\leq \frac{V^2(V+2A^{(0)})^2}{(V+A^{(0)})^4}\left\{n(\bar{\theta}^{(0)}-\bar{Y})^2 + n\left[\frac{\Delta}{n-1} - (A^{(0)}+V)\right]^2\right\} + \mathcal{O}(1) \\ &= \left[\left(\frac{V^2+2VA^{(0)}}{V^2+2VA^{(0)}+(A^{(0)})^2}\right)^2\right] f(x^{(0)}) + \mathcal{O}(1).\end{aligned} \tag{2.76}$$

2.3.4 Proof of Lemma 2.2.4

Throughout the proof, we write $f_n(x)$ as $f(x)$ for simplicity. Recall that the small set is defined by $R = \{x \in \mathcal{X} : f(x) \leq d\}$ where $d > 2b/(1-\lambda_T)$ and $x = (\mu, A, \theta_1, \ldots, \theta_n)$. When $b = \mathcal{O}(1)$ and $\lambda_T = \Theta(1)$, we can choose $d = \mathcal{O}(1)$. Our goal is to show the minorization volume ϵ satisfying

$$P(x, \cdot) \geq \epsilon Q(\cdot), \quad \forall x \in R, \tag{2.77}$$

is asymptotically bounded away from 0. Denoting $\hat{A} := \frac{\Delta}{n-1} - V$, we have

$$\begin{aligned}R &= \left\{x \in \mathcal{X} : n(\bar{\theta}-\bar{Y})^2 + n\left[\left(\frac{\Delta}{n-1} - V\right) - A\right]^2 \leq d\right\} \\ &\subseteq \left\{x \in \mathcal{X} : |\bar{\theta}-\bar{Y}| \leq \sqrt{\frac{d}{n}}\right\} \bigcap \left\{x \in \mathcal{X} : |A-\hat{A}| \leq \sqrt{\frac{d}{n}}\right\}\end{aligned} \tag{2.78}$$

Denoting

$$R' := \left\{x \in \mathcal{X} : |\bar{\theta}-\bar{Y}| \leq \sqrt{\frac{d}{n}}, |A-\hat{A}| \leq \sqrt{\frac{d}{n}}\right\} \tag{2.79}$$

since $R \subseteq R'$, it suffices to show the minorization volume ϵ satisfying

$$P(x^{(0)}, \cdot) \geq \epsilon Q(\cdot), \quad \forall x^{(0)} \in R', \tag{2.80}$$

is asymptotically bounded away from 0. One common technique to obtain ϵ is by integrating the infimum of densities of $P(x^{(0)}, \cdot)$ where in our case the infimum is over all $\bar{\theta}^{(0)}$ and $A^{(0)}$ such that $|\bar{\theta}^{(0)} - \bar{Y}| \leq \sqrt{\frac{d}{n}}$ and $|A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}$.

Note that the intuition behind the proof is: since R' is determined by $|\bar{\theta}^{(0)} - \bar{Y}| \leq \sqrt{\frac{d}{n}}$ and $|A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}$. The size of uncertainties of the initial $\bar{\theta}^{(0)}$ and $A^{(0)}$ is of order $\mathcal{O}(1/\sqrt{n})$. Therefore, for any fixed initial state $x^{(0)} \in R'$, if the transition kernel $P(x^{(0)}, \cdot)$ concentrates at a rate of $\Omega(1/\sqrt{n})$ then ϵ is bounded away from 0.

For the density function of the Markov transition kernel $P(x^{(0)}, \cdot)$, recall the order of Gibbs sampler

$$\begin{aligned}
\mu^{(1)} &\sim \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}\right), \\
\theta_i^{(1)} &\sim \mathcal{N}\left(\frac{\mu^{(1)} V + Y_i A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)} V}{V + A^{(0)}}\right), \quad i = 1, \ldots, n \\
A^{(1)} &\sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n} (\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right).
\end{aligned} \tag{2.81}$$

Then ϵ can be computed using the three steps of integration according to the reverse order of the Gibbs sampler:

1. For given $\mu^{(1)}$ and $\{\theta_i^{(1)}\}$, integrating the infimum of the density of $A^{(1)}$. Note that the infimum is over a subset of $\bar{\theta}^{(0)}$ and $A^{(0)}$. However,

$$A^{(1)} \sim \mathbf{IG}\left(a + \frac{n-1}{2}, b + \frac{1}{2}\sum_{i=1}^{n} (\theta_i^{(1)} - \bar{\theta}^{(1)})^2\right) \tag{2.82}$$

   does not depend on $\bar{\theta}^{(0)}$ and $A^{(0)}$. Therefore, the integration of the infimum of the density in this step always equals one;

2. For given $\mu^{(1)}$, integrating the infimum of the densities of $\{\theta_i^{(1)}\}$. We first note that $\{\theta_i^{(1)}\}$ appear in the densities only in the forms of $\bar{\theta}^{(1)}$ and $S = \frac{\sum_i (\theta_i^{(1)} - \bar{\theta}^{(1)})^2}{n-1}$. Therefore, instead of integrating over $(\theta_1^{(1)}, \ldots, \theta_n^{(1)})$ we can integrate over $\bar{\theta}^{(1)}$ and S. Furthermore, we have shown $\bar{\theta}^{(1)}$ is conditional independent with S given $A^{(0)}$ in the proof of Lemma 2.3.3, we can integrate them separately. Finally, we note that the infimum is over $\left\{A^{(0)} : |A^{(0)} - \hat{A}| \leq \sqrt{\frac{d}{n}}\right\}$. Overall, we need to show

$\tilde{g}_n(\mu^{(1)})$ is lower bounded away from 0, which is defined by

$$\begin{aligned}\tilde{g}_n(\mu^{(1)}) &:= \int \mathrm{d}S\mathrm{d}\bar{\theta} \inf_{x^{(0)}\in R'} \left\{ f_S(A^{(0)}, n; S)\mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right)\right\} \\ &\geq \left[\int \mathrm{d}S \inf_{x^{(0)}\in R'} f_S(A^{(0)}, n; S)\right] \\ &\quad \cdot \left[\int \mathrm{d}\bar{\theta} \inf_{x^{(0)}\in R'} \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right)\right],\end{aligned} \tag{2.83}$$

where $f_S(A^{(0)}, n; S)$ denotes the density function of $S = \frac{\sum_i(\theta_i - \bar{\theta})^2}{n-1}$ for given $A^{(0)}$, with

$$\theta_i \sim \mathcal{N}\left(\frac{\mu^{(1)}V + Y_iA^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{V + A^{(0)}}\right), \quad i = 1, \ldots, n, \tag{2.84}$$

and $\mathcal{N}\left(\frac{\mu^{(1)}V+\bar{Y}A^{(0)}}{V+A^{(0)}}, \frac{A^{(0)}V}{n(V+A^{(0)})}; \bar{\theta}\right)$ denotes the density function of

$$\bar{\theta} \sim \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}\right). \tag{2.85}$$

3. Finally, we integrate the infimum of the densities of $\mu^{(1)}$ to get ϵ. That is,

$$\epsilon = \int \mathrm{d}\mu \left\{ \tilde{g}_n(\mu) \inf_{x^{(0)}\in R'} \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}; \mu\right)\right\}. \tag{2.86}$$

In the following, we show ϵ is lower bounded away from 0 in three steps.

First, it is easy to see that the density of S does not depend on $\mu^{(1)}$. We show

$$\int \mathrm{d}S \inf_{x^{(0)}\in R'} f_S(A^{(0)}, n; S) = \Theta(1). \tag{2.87}$$

Second, we show

$$\int \mathrm{d}\bar{\theta} \inf_{x^{(0)}\in R'} \mathcal{N}\left(\frac{\mu^{(1)}V + \bar{Y}A^{(0)}}{V + A^{(0)}}, \frac{A^{(0)}V}{n(V + A^{(0)})}; \bar{\theta}\right) \geq 1 - \mathrm{erf}\left(\frac{C|\mu| + C'}{\sqrt{2}}\right) \tag{2.88}$$

where $\mathrm{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\mathrm{d}t$ and C and C' are some constants.

Finally, we complete the proof by showing

$$\int \mathrm{d}\mu \left\{ \left(1 - \mathrm{erf}(\frac{C|\mu| + C'}{\sqrt{2}})\right) \inf_{x^{(0)}\in R'} \mathcal{N}\left(\bar{\theta}^{(0)}, \frac{A^{(0)}}{n}; \mu\right)\right\} = \Theta(1). \tag{2.89}$$

Proof of Eq. (2.87)

We omit the superscripts for simplicity. That is, we show

$$\int \mathrm{d}S \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} f_S(A, n; S) = \Theta(1). \tag{2.90}$$

Following the proof of Lemma 2.3.3 from Eq. (3.136) to Eq. (3.196), defining

$$\eta_i := \theta_i - \frac{Y_i A}{V+A} \sim \mathcal{N}\left(\frac{\mu V}{V+A}, \frac{AV}{V+A}\right), \tag{2.91}$$

we know

$$\mathbb{E}\left[\left|S - \frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1} - \left(\frac{A}{V+A}\right)^2 \frac{\Delta}{n-1}\right|^2\right] = \mathcal{O}(1/n). \tag{2.92}$$

Therefore, defining

$$S' := \frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1} + \left(\frac{A}{V+A}\right)^2 \frac{\Delta}{n-1} \tag{2.93}$$

and denoting $f'_{S'}(A,n;S')$ as the density of S', it suffices to show

$$\int \mathrm{d}S' \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f'_{S'}(A,n;S') = \Theta(1). \tag{2.94}$$

Furthermore, note that under $|A-\hat{A}| \le \sqrt{\frac{d}{n}}$, we have $\frac{V+A}{AV} = \frac{V+\hat{A}}{\hat{A}V} + \mathcal{O}(1/\sqrt{n}) = \Theta(1)$. Then it suffices to show

$$\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A,n;S'') = \Theta(1), \tag{2.95}$$

where

$$S'' := \frac{V+A}{AV}S' = \frac{V+A}{AV}\frac{\sum_i(\eta_i-\bar{\eta})^2}{n-1} + \frac{1}{V}\left(\frac{A}{V+A}\right)\frac{\Delta}{n-1} \tag{2.96}$$

and $f''_{S''}(A,n;S'')$ is the density function of S''.

Next, note that $\frac{V+A}{AV}\sum_i(\eta_i-\bar{\eta})^2 \sim \chi^2_{n-1}$, we have

$$\frac{\frac{V+A}{AV}\sum_i(\eta_i-\bar{\eta})^2 - (n-1)}{\sqrt{2(n-1)}} \xrightarrow{d} \mathcal{N}(0,1), \tag{2.97}$$

which does not depend on n. We define $\tilde{f}(z,A;x), \forall z \in \mathbb{R}$ as the density function of a random variable

$$\tilde{X}_{z,A} := z + \frac{\frac{V+A}{AV}\sum_i(\eta_i-\bar{\eta})^2 - (n-1)}{\sqrt{2(n-1)}}, \tag{2.98}$$

then we know $\tilde{X}_{z,A} \xrightarrow{d} \mathcal{N}(z,1)$.

The rest of the proof is first to lower bound $\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\le\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A,n;S'')$ using the density function $\tilde{f}(z,A;x)$ and then show it is asymptotically lower bounded away from 0.

Notice that $\frac{1}{V}\left(\frac{A}{V+A}\right)\frac{\Delta}{n-1}$ is not random, and there exists a constant C_0 such that

$$\left(\max_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\frac{A}{V+A}-\min_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\frac{A}{V+A}\right)\frac{\Delta/V}{n-1}\leq\frac{C_0}{\sqrt{n-1}}. \tag{2.99}$$

Finally we have

$$\begin{aligned}
&\int \mathrm{d}S'' \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} f''_{S''}(A,n;S'')\\
&\geq \inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\int \mathrm{d}x \min\left\{\tilde{f}\left(-\frac{C_0}{\sqrt{2}},A;x\right),\tilde{f}\left(+\frac{C_0}{\sqrt{2}},A;x\right)\right\}\\
&=1-\sup_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\int_{-\sqrt{2}C_0}^{\sqrt{2}C_0}\mathrm{d}x\tilde{f}(0,A;x)\\
&=1-\sup_{\{A:|A-\hat{A}|\leq\sqrt{d/n}\}}\mathbb{P}(-\sqrt{2}C_0\leq\tilde{X}_{0,A}\leq\sqrt{2}C_0)\\
&\to 1-\int_{-\sqrt{2}C_0}^{\sqrt{2}C_0}\mathrm{d}x\mathcal{N}(0,1;x)=\Theta(1).
\end{aligned} \tag{2.100}$$

Proof of Eq. (2.88)

We again omit the subscripts for simplicity. The goal is to lower bound

$$\int \mathrm{d}\bar{\theta}\inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\mathcal{N}\left(\frac{\mu V+\bar{Y}A}{V+A},\frac{AV}{n(V+A)};\bar{\theta}\right) \tag{2.101}$$

Note that there exists some constants C_1 and C_2 such that

$$\max_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{\mu V+\bar{Y}A}{V+A}-\min_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{\mu V+\bar{Y}A}{V+A}\leq\frac{C_1|\mu|+C_2}{\sqrt{n}}, \tag{2.102}$$

and another constant C_3 such that

$$\min_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\frac{AV}{n(V+A)}\geq\frac{C_3}{n}. \tag{2.103}$$

Therefore, we have

$$\begin{aligned}
&\int \mathrm{d}\bar{\theta}\inf_{\left\{A:|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}}\mathcal{N}\left(\frac{\mu V+\bar{Y}A}{V+A},\frac{AV}{n(V+A)};\bar{\theta}\right)\\
&\geq 2\int_{(C_1|\mu|+C_2)/\sqrt{n}}^{\infty}\mathrm{d}x\,\mathcal{N}(0,C_3/n;x)\\
&=2\int_{C_4|\mu|+C_5}^{\infty}\mathrm{d}x\,\mathcal{N}(0,1;x)\\
&=1-\mathrm{erf}\left(\frac{C_4|\mu|+C_5}{\sqrt{2}}\right),
\end{aligned} \tag{2.104}$$

where $C_4 := \frac{C_1}{\sqrt{C_3}}$ and $C_5 := \frac{C_2}{\sqrt{C_3}}$.

Proof of Eq. (2.89)

We omit the subscripts for simplicity. We show the following is asymptotically bounded away from 0:

$$\int \mathrm{d}\mu \left\{ \left(1 - \mathrm{erf}\left(\frac{C_4|\mu| + C_5}{\sqrt{2}}\right)\right) \inf_{x \in R'} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \tag{2.105}$$

Note that there exists $A'_n \in [\hat{A} - \sqrt{d/n}, \hat{A} + \sqrt{d/n}]$ such that

$$\begin{aligned} &\inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\leq\sqrt{\frac{d}{n}},|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \\ &\quad = \min\left\{ \mathcal{N}\left(\bar{Y} - \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right), \mathcal{N}\left(\bar{Y} + \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \right\} \end{aligned} \tag{2.106}$$

Therefore, we have

$$\begin{aligned} &\int_{-\infty}^{\infty} \mathrm{d}\mu \left\{ \left(1 - \mathrm{erf}\left(\frac{C_4|\mu| + C_5}{\sqrt{2}}\right)\right) \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\leq\sqrt{\frac{d}{n}},|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\ &\geq \int_0^{2\bar{Y}} \mathrm{d}\mu \left\{ \left(1 - \mathrm{erf}\left(\frac{C_4|\mu| + C_5}{\sqrt{2}}\right)\right) \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\leq\sqrt{\frac{d}{n}},|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \right\} \\ &\geq \left(1 - \mathrm{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \int_0^{2\bar{Y}} \mathrm{d}\mu \inf_{\left\{(\bar{\theta},A):|\bar{\theta}-\bar{Y}|\leq\sqrt{\frac{d}{n}},|A-\hat{A}|\leq\sqrt{\frac{d}{n}}\right\}} \mathcal{N}\left(\bar{\theta}, \frac{A}{n}; \mu\right) \\ &= \left(1 - \mathrm{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \\ &\quad \cdot \left[\int_0^{\bar{Y}} \mathrm{d}\mu \mathcal{N}\left(\bar{Y} + \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{\bar{Y}}^{2\bar{Y}} \mathrm{d}\mu \mathcal{N}\left(\bar{Y} - \sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \right] \\ &= \left(1 - \mathrm{erf}\left(\frac{C_4|2\bar{Y}| + C_5}{\sqrt{2}}\right)\right) \\ &\quad \cdot \left[\int_{-\bar{Y}}^{0} \mathrm{d}\mu \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_0^{\bar{Y}} \mathrm{d}\mu \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \right] \end{aligned} \tag{2.107}$$

Finally, we show

$$\int_{-\bar{Y}}^{0} \mathrm{d}\mu \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_0^{\bar{Y}} \mathrm{d}\mu \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \tag{2.108}$$

is asymptotically bounded away from 0. Note that when $n \to \infty$, we have $A'_n \to \hat{A}$. So the density

functions $\mathcal{N}\left(\pm\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right)$ concentrate on 0. Therefore

$$\begin{aligned}
&\int_{-\bar{Y}}^{0} \mathrm{d}\mu \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) + \int_{0}^{\bar{Y}} \mathrm{d}\mu \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{A'_n}{n}; \mu\right) \\
&\to \int_{-\infty}^{0} \mathrm{d}\mu \mathcal{N}\left(\sqrt{\frac{d}{n}}, \frac{\hat{A}}{n}; \mu\right) + \int_{0}^{\infty} \mathrm{d}\mu \mathcal{N}\left(-\sqrt{\frac{d}{n}}, \frac{\hat{A}}{n}; \mu\right) \\
&= 1 - \int_{-\sqrt{d/n}}^{\sqrt{d/n}} \mathrm{d}x \mathcal{N}\left(0, \frac{\hat{A}}{n}; x\right) \\
&= 1 - \int_{-\sqrt{d}}^{\sqrt{d}} \mathrm{d}x \mathcal{N}(0, \hat{A}; x) = \Theta(1).
\end{aligned} \tag{2.109}$$

2.3.5 Proof of Lemma 2.2.5

In this proof, we write $f_n(x)$ as $f(x)$ for simplicity. We first consider a Markov chain starting from initial state $x^{(0)}$ defined by Eq. (2.15). By Eq. (2.14), we have $A^{(0)} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} - V$ for large enough n, which implies $f(x^{(0)}) = 0$. Therefore, for large enough n, we have $\mathbb{E}(f(x^{(1)})) \leq b$ from Lemma 2.2.3. Furthermore, we can continue to get upper bounds $\mathbb{E}(f(x^{(i)})) \leq ib$ for all $i = 1, \ldots, k$. This implies

$$\mathbb{E}\left[\left(\left(\frac{\Delta}{n-1} - V\right) - A^{(i)}\right)^2\right] \leq i\frac{b}{n}, \quad i = 1, \ldots, k. \tag{2.110}$$

By the Markov's inequality, we have

$$\mathbb{P}\left(\left|A^{(i)} - \left(\frac{\Delta}{n-1} - V\right)\right| \geq \left|T - \left(\frac{\Delta}{n-1} - V\right)\right|\right) \leq \frac{i}{n}\frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}, \tag{2.111}$$

for $i = 1, \ldots, k$. Therefore, we have

$$\sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \leq \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2} \sum_{i=1}^{k} \frac{i}{n} = \frac{k(1+k)}{2n} \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}. \tag{2.112}$$

Next, we consider a Markov chain starting from π. According to Lemma 2.2.3, we have

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(1 - \left(\frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)^2\right) f(x)\right] \\
&= \mathbb{E}_\pi\left[\left(1 + \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)\left(1 - \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right) f(x)\right] \\
&= \mathbb{E}_\pi\left[\left(1 + \frac{V^2 + 2VA}{V^2 + 2VA + A^2}\right)\left(\frac{A}{V + A}\right)^2 f(x)\right] \leq b,
\end{aligned} \tag{2.113}$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expectation is over $x \sim \pi(\cdot)$. Note that by Hölder's inequality (in the reverse

way)

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(1+\frac{V^2+2VA}{V^2+2VA+A^2}\right)\left(\frac{A}{V+A}\right)^2 f(x)\right] \\
&\geq \mathbb{E}_\pi\left[\left(\frac{A}{V+A}\right)^2 f(x)\right] \\
&\geq [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2\left\{\mathbb{E}_\pi\left[\left(\frac{A}{V+A}\right)^{-2}\right]\right\}^{-1} \\
&= [\mathbb{E}_\pi(f(x)^{\frac{1}{2}})]^2/\mathbb{E}_\pi[(1+V/A)^2].
\end{aligned} \tag{2.114}$$

Therefore, we have

$$\mathbb{E}_\pi(f(x)^{\frac{1}{2}}) \leq \sqrt{b}\sqrt{1+2V\mathbb{E}_\pi(1/A)+V^2\mathbb{E}_\pi(1/A^2)}. \tag{2.115}$$

Next, we show $\mathbb{E}_\pi(1/A) \leq 2/\delta$ and $\mathbb{E}_\pi(1/A^2) \leq 2/\delta^2$ for large enough n.

Lemma 2.3.4. *There exists a positive integer N, which only depends on a, b, V, and δ, such that for all $n \geq N$, we have*

$$\mathbb{E}_\pi(1/A) \leq 2/\delta, \quad \mathbb{E}_\pi(1/A^2) \leq 2/\delta^2. \tag{2.116}$$

Proof. The posterior distribution can be written as

$$\pi(x \,|\, Y_1, \ldots, Y_n) = \frac{f_a(x, Y_1, \ldots, Y_n)}{\int f_a(x, Y_1, \ldots, Y_n)\mathrm{d}x}, \tag{2.117}$$

where we use $f_a(x, Y_1, \ldots, Y_n)$ to denote the joint distribution of x and $\{Y_i\}$ when $\mathbf{IG}(a, b)$ is used as the prior for A. That is,

$$\begin{aligned}
&f_a(x, Y_1, \ldots, Y_n) \\
&= \frac{b^a}{\Gamma(a)}A^{-a-1}e^{-b/A}\prod_{i=1}^n \frac{1}{\sqrt{2\pi A}}e^{-\frac{(\theta_i-\mu)^2}{2A}}\frac{1}{\sqrt{2\pi}}e^{-\frac{(Y_i-\theta_i)^2}{2V}} \\
&= \frac{1}{(2\pi)^n}\frac{b^a}{\Gamma(a)}A^{-a-1-\frac{n}{2}}e^{-b/A}\exp\left[-\sum_{i=1}^n\left(\frac{(\theta_i-\mu)^2}{2A}+\frac{(Y_i-\theta_i)^2}{2V}\right)\right].
\end{aligned} \tag{2.118}$$

Now using $\frac{1}{A}f_a(x, Y_1, \ldots, Y_n) = \frac{a}{b}f_{a+1}(x, Y_1, \ldots, Y_n)$, we have

$$\mathbb{E}_\pi(1/A) = \frac{a}{b}\frac{\int f_{a+1}(x, Y_1, \ldots, Y_n)\mathrm{d}x}{\int f_a(x, Y_1, \ldots, Y_n)\mathrm{d}x}, \quad \mathbb{E}_\pi(1/A^2) = \frac{a^2}{b^2}\frac{\int f_{a+2}(x, Y_1, \ldots, Y_n)\mathrm{d}x}{\int f_a(x, Y_1, \ldots, Y_n)\mathrm{d}x}. \tag{2.119}$$

Therefore, it suffices to show the ratios $\frac{\int f_{a+1}(x,Y_1,\ldots,Y_n)\mathrm{d}x}{\int f_a(x,Y_1,\ldots,Y_n)\mathrm{d}x}$ and $\frac{\int f_{a+2}(x,Y_1,\ldots,Y_n)\mathrm{d}x}{\int f_a(x,Y_1,\ldots,Y_n)\mathrm{d}x}$ are (asymptotically) bounded. Next, we focus on the first ratio. The second ratio can be proved using a similar argument.

Using the fact that

$$\begin{aligned}
&\int \exp\left[-\left(\frac{V(\theta_i-\mu)^2+A(Y_i-\theta_i)^2}{2AV}\right)\right]\mathrm{d}\theta_i \\
&= \left(\int \exp\left[-\frac{\left(\theta-\frac{V\mu+YA}{A+V}\right)^2}{\frac{2AV}{A+V}}\right]\mathrm{d}\theta_i\right)\left(\exp\left[-\frac{(Y_i-\mu)^2}{2(V+A)}\right]\right) \\
&= \sqrt{2\pi\frac{2AV}{V+A}}\exp\left[-\frac{(Y_i-\mu)^2}{2(V+A)}\right],
\end{aligned} \tag{2.120}$$

and

$$\begin{aligned}
&\int \exp\left[-\frac{\sum_{i=1}^n(Y_i-\mu)^2}{2(V+A)}\right]\mathrm{d}\mu \\
&= \left(\int \exp\left[-\frac{(\mu-\bar{Y})^2}{2(V+A)/n}\right]\mathrm{d}\mu\right)\left(\exp\left[-\frac{\sum_i Y_i^2-n\bar{Y}^2}{2(V+A)}\right]\right) \\
&= \exp\left[-\frac{\sum_{i=1}^n(Y_i-\bar{Y})^2}{2(V+A)}\right]\sqrt{2\pi\frac{2(V+A)}{n}},
\end{aligned} \tag{2.121}$$

we can write $\mathbb{E}_\pi(1/A)$ as a function of $\Delta=\sum_i(Y_i-\bar{Y})^2$. Denote $h_n(\Delta):=\mathbb{E}_\pi(1/A)$, then we have

$$h_n(\Delta):=\frac{\int A^{-a-2}e^{-b/A}(V+A)^{\frac{-n+1}{2}}\exp\left[-\frac{\Delta}{2(V+A)}\right]\mathrm{d}A}{\int A^{-a-1}e^{-b/A}(V+A)^{\frac{-n+1}{2}}\exp\left[-\frac{\Delta}{2(V+A)}\right]\mathrm{d}A}. \tag{2.122}$$

Next, we show $h_n((n-1)(c+V))$ is (asymptotically) bounded for any fixed $c>0$. Note that

$$\begin{aligned}
&\int A^{-a-1}e^{-b/A}(V+A)^{\frac{-n+1}{2}}\exp\left[-\frac{\Delta}{2(V+A)}\right]\mathrm{d}A \\
&\quad= \int A^{-a-1}e^{-b/A}\left\{\frac{1}{\sqrt{V+A}}\exp\left[-\frac{\frac{\Delta}{n-1}}{2(V+A)}\right]\right\}^{n-1}\mathrm{d}A.
\end{aligned} \tag{2.123}$$

We change variable $y=\frac{1}{\sqrt{V+A}}$ and apply the Laplace approximation. Note that for any $c>0$, let $y_0=\arg\max_y\left[y\exp\left(-\frac{c+V}{2}y^2\right)\right]$, then $y_0=\frac{1}{\sqrt{c+V}}$. Therefore, by the Laplace approximation [ZC04, Thm. 1, Chp. 19.2.4], we have

$$\begin{aligned}
h_n((n-1)(c+V)) &= \frac{c^{-a-2}e^{-b/c}\left[y_0\exp\left(-\frac{c+V}{2}y_0^2\right)\right]^{n-1}(1+\mathcal{O}(n^{-\frac{1}{2}}))}{c^{-a-1}e^{-b/c}\left[y_0\exp\left(-\frac{c+V}{2}y_0^2\right)\right]^{n-1}(1+\mathcal{O}(n^{-\frac{1}{2}}))} \\
&= \frac{1}{c}(1+\mathcal{O}(n^{-1/2})),
\end{aligned} \tag{2.124}$$

where the term $\mathcal{O}((n^{-1/2})$ only depends on constants a, b, and V. Finally, since for all $n\geq N_0$ we have $\Delta\geq(n-1)(V+\delta)$, this implies $h_n(\Delta)\leq\frac{1}{\delta}(1+\mathcal{O}(n^{-1/2})),\forall n\geq N_0$. Therefore, there exists large enough positive integer N_0, which only depends on a, b, V, and δ, such that for all $n\geq N_0$, we have $\mathbb{E}_\pi(1/A)=h_n(\Delta)\leq\frac{1}{\delta}(1+\mathcal{O}(n^{-1/2}))\leq\frac{2}{\delta}$.

For $\mathbb{E}_\pi(1/A^2)$, we can follow a similar argument to show that $\mathbb{E}_\pi(1/A^2)\leq\frac{2}{\delta^2}$ for large enough n. Therefore, we can conclude that there exists large enough positive integer N, which only depends on a, b,

V, and δ, such that for all $n \geq N$, we have both $\mathbb{E}_\pi(1/A) \leq \frac{2}{\delta}$ and $\mathbb{E}_\pi(1/A^2) \leq \frac{2}{\delta^2}$. □

By Lemma 2.3.4, we have $\sqrt{1 + 2V\mathbb{E}_\pi(1/A) + V^2\mathbb{E}_\pi(1/A^2)} \leq 1 + 2V/\delta$ for large enough n. Therefore, we get

$$\mathbb{E}_\pi\left(\left|\left(\frac{\Delta}{n-1} - V\right) - A\right|\right) \leq \sqrt{\frac{b}{n}}(2V/\delta + 1). \tag{2.125}$$

Thus, by the Markov's inequality

$$\begin{aligned}\pi(R_T^c) &= \mathbb{P}_\pi\left(\left|\left(\frac{\Delta}{n-1} - V\right) - A\right| \geq \left|\left(\frac{\Delta}{n-1} - V\right) - T\right|\right) \\ &\leq \frac{\sqrt{\frac{b}{n}}(2V/\delta + 1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|}.\end{aligned} \tag{2.126}$$

Finally, we have

$$\begin{aligned}&k\,\pi(R_T^c) + \sum_{i=1}^{k} P^i(x^{(0)}, R_T^c) \\ &\leq \frac{k}{\sqrt{n}} \frac{\sqrt{b}(2V/\delta + 1)}{\left|\left(\frac{\Delta}{n-1} - V\right) - T\right|} + \frac{k(1+k)}{2n} \frac{b}{\left[T - \left(\frac{\Delta}{n-1} - V\right)\right]^2}.\end{aligned} \tag{2.127}$$

Chapter 3

Optimal scaling of random-walk Metropolis algorithms on general target distributions

One main limitation of the existing optimal scaling results for Metropolis–Hastings algorithms is that the assumptions on the target distribution are unrealistic. In this chapter, we consider optimal scaling of random-walk Metropolis algorithms on general target distributions in high dimensions arising from practical MCMC models from Bayesian statistics. For optimal scaling by maximizing expected squared jumping distance (ESJD), we show the asymptotically optimal acceptance rate 0.234 can be obtained under general realistic sufficient conditions on the target distribution. The new sufficient conditions are easy to be verified and may hold for some general classes of MCMC models arising from Bayesian statistics applications, which substantially generalize the product i.i.d. condition required in most existing literature of optimal scaling. Furthermore, we show one-dimensional diffusion limits can be obtained under slightly stronger conditions, which still allow dependent coordinates of the target distribution. We also connect the new diffusion limit results to complexity bounds of Metropolis algorithms in high dimensions.

3.1 Background on optimal scaling

Practical implementations of Metropolis–Hastings algorithms suffer from slow mixing for at least two reasons: the Markov chain moves very slowly to the target distribution when the proposed jumps are too short; the Markov chain stays at a state for most of the time when the proposed jumps are long but the chain ends up in low probability areas of the target distribution. The optimal scaling problem [RGG97] considers the choice of proposed distribution to optimize mixing of the Metropolis–Hastings algorithm. We focus on one of the most popular MCMC algorithms, the RWM algorithm. This algorithm proceeds by running a Markov chain $\{X^d(t), t = 0, \ldots, \infty\}$ as follows. Given a target distribution π^d on the state space $\mathbb{R}^d$ and the current state $X^d(t) = x^d$, a new state is proposed by $Y^d \sim \mathcal{N}(x^d, \sigma_d^2 I)$, which is sampled from a multivariate Gaussian distribution centered at x^d, then the proposal is accepted with probability $\min\{1, \pi^d(Y^d)/\pi^d(x^d)\}$ so that $X^d(t+1) = Y^d$. Otherwise the proposal is rejected and

$X^d(t+1) = x^d$. This is precisely to ensure the Markov chain is reversible with respect to the target distribution π^d. It can be shown that the normal proposals automatically make the RWM algorithm π^d-irreducible, aperiodic, and hence ergodic [RS94; MT96]. Therefore, it will converge asymptotically to π^d in law. Note that the only computational cost involved in calculating the acceptance probabilities is the relative ratio of densities. Within the class of all Metropolis–Hastings algorithms, the RWM algorithm is still widely used in many applications because of its simplicity and robustness.

3.1.1 Optimal scaling via diffusion limits

The most common technique to prove optimal scaling results is to show a weak convergence to diffusion limits as the dimension of a sequence of target densities converges to infinity [RGG97; RR98]. More specifically, even though different coordinates of the Markov chain are *not* independent *nor* even individually Markovian, when the proposal is appropriately scaled according to the dimension, the sequence of sped-up stochastic processes formed by one fixed coordinate of each Markov chain converges to an appropriate Markovian Langevin diffusion process. The limiting diffusion limit admits a straightforward efficiency maximization problem which leads to asymptotically optimal acceptance rate of the proposed moves for the Metropolis–Hastings algorithm. In [RGG97], the target distribution π^d is assumed to be an d-dimensional product density with respect to Lebesgue measure, that is

$$\pi^d(x^d) = \prod_{i=1}^{d} f(x_i), \tag{3.1}$$

where $x^d = (x_1, x_2, \dots, x_d)$. It is shown that with the choice of scaling $\sigma_d^2 = \ell^2/(d-1)$ for some fixed $\ell > 0$, individual components of the resulting Markov chain converge to the solution of a stochastic differential equation (SDE). More specifically, denoting $X^d = (X_1^d, X_2^d, \dots, X_d^d)$, the first coordinate of the RWM algorithm, X_1^d, sped up by a factor of d, i.e. $\{X_1^d(\lfloor dt \rfloor), t = 0, 1, \dots\}$, converges weakly in the usual Skorokhod topology to a limiting ergodic Langevin diffusion.

Proposition 3.1.1. *[RGG97, Theorem 1.1] Suppose density f satisfies that f'/f is Lipschitz continuous and*

$$\int \left[\frac{f'(x)}{f(x)}\right]^8 f(x)\mathrm{d}x < \infty, \quad \int \left[\frac{f''(x)}{f(x)}\right]^4 f(x)\mathrm{d}x < \infty. \tag{3.2}$$

Then for $U^d(t) := X_1^d(\lfloor dt \rfloor)$, as $d \to \infty$, we have $U^d \Rightarrow U$, where $\Rightarrow$ denotes weak convergence in Skorokhod topology, and U satisfies the following Langevin SDE

$$\mathrm{d}U(t) = (h(\ell))^{1/2}\mathrm{d}B(t) + h(\ell)\frac{f'(U(t))}{2f(U(t))}\mathrm{d}t, \tag{3.3}$$

with $h(\ell) := 2\ell^2\Phi(-\ell\sqrt{\tilde{I}}/2)$ is the speed measure for the diffusion process, $\tilde{I} := \int \left[\frac{f'(x)}{f(x)}\right]^2 f(x)\mathrm{d}x$, and Φ being the standard Gaussian cumulative density function.

This weak convergence result leads to the interpretation that, started in stationarity and applied to target measures of the i.i.d. form, the RWM algorithm will take on the order of d steps to explore the invariant measure. Furthermore, it may be shown that the value of ℓ which maximizes the speed measure $h(\ell)$ and, therefore, maximizes the speed of convergence of the limiting diffusion, leads to a universal

acceptance probability, for the RWM algorithm applied to targets of i.i.d. forms, of approximately 0.234. Proposition 3.1.1 is proved in [RGG97] using the generator approach [EK86]. The same method of proof has also been applied to derive optimal scaling results for other types of MCMC algorithms: for example, the convergence of MALA to diffusion limits when $\sigma_d^2 = \ell^2/d^{1/3}$ (see e.g. [RR98; RR01; BPS04; CRR05; NR06]) with asymptotically optimal acceptance rate 0.574.

3.1.2 Optimal scaling by maximizing ESJD

Another popular technique to prove optimal scaling is by maximizing expected squared jumping distance (ESJD) [PG10; ARR11; RR14], which is defined as follows.

Definition 3.1.2. (Expected Squared Jumping Distance)

$$\mathrm{ESJD}(d) := \mathbb{E}_{X^d \sim \pi^d} \mathbb{E}_{Y^d} \left[\|Y^d - X^d\|^2 \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right] \tag{3.4}$$

where the expectation over Y^d is taken for $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1} I)$ for given $X^d = x^d$, and $\|\cdot\|$ denotes the Euclidean distance, i.e. $\|Y^d - X^d\|^2 = \sum_{i=1}^d (Y_i - X_i)^2$.

Choosing a proposal variance to maximize ESJD is equivalent to minimizing the first-order auto-correlation of the Markov chain, and thus maximizing the efficiency if the higher order auto-correlations are monotonically increasing with respect to the first-order auto-correlation [PG10]. Furthermore, if weak convergence to a diffusion limit is established, then the ESJD converges to the quadratic variation of the diffusion limit. This suggests that maximizing the ESJD is a reasonable problem. For example, Atchadé, Roberts, and Rosenthal [ARR11] considered to maximize the ESJD to choose optimal temperature spacings for Metropolis-coupled Markov chain Monte Carlo and simulated tempering algorithms. Later, Roberts and Rosenthal [RR14] proved a diffusion limit for the simulated tempering algorithms. Using a new comparison of asymptotic variance of diffusions, Roberts and Rosenthal [RR14] showed the results in the choice of temperatures in [ARR11] does indeed minimize the asymptotic variance of all functionals. Another example is the optimal scaling result for HMC, with asymptotically optimal acceptance rate 0.651 with the choice of scaling $\sigma_d^2 = \ell^2/d^{1/4}$ for some fixed $\ell > 0$ [Bes+13], is proven by maximizing the ESJD.

Although establishing weak convergence of diffusion limits gives stronger guarantee than maximizing ESJD, the price to pay is to require stronger conditions on the target distribution. Maximizing ESJD instead can lead to (much) weaker conditions on the target distribution. Later in this chapter, we will show that we are able to relax the restrictive product i.i.d. condition on the target distribution for both cases. In particular, the new sufficient conditions on the target distribution for maximizing ESJD are weak enough to allow target distributions arising from realistic MCMC models.

3.1.3 Background on complexity bounds

Because of the big data world, in recent years, there is much interest in the "large d, large n" or "large d, small n" high-dimensional regime, where d is the number of parameters and n is the sample size. Rajaratnam and Sparks [RS15] use the term convergence complexity to denote the ability of a high-dimensional MCMC scheme to draw samples from the posterior, and how the ability to do so changes as the dimension of the parameter set grows. This requires the study of computer-science-style complexity

bounds [Cob65; Coo71] in terms of running time complexity order as the "size" of the problem goes to infinity. In the Markov chain context, computer scientists have been bounding convergence times of Markov chain algorithms focusing largely on spectral gap bounds for Markov chains [SJ89; LV03; Vem05; LV06; WSH09a; WSH09b]. In contrast, statisticians usually study total variation distance or other metric for MCMC algorithms. In order to bridge the gap between statistics-style convergence bounds, and computer-science-style complexity results, in one direction, Yang and Rosenthal [YR17] recently show that complexity bounds for MCMC can be obtained by quantitative bounds using a modified drift-and-minorization approach. In another direction, Roberts and Rosenthal [RR16] connect existing results on diffusion limits of MCMC algorithm to the computer science notion of algorithm complexity. The main result in [RR16] states that any weak limit of a Markov process implies a corresponding complexity bound in an appropriate metric. More specifically, Roberts and Rosenthal [RR16] connect the diffusion limits to complexity bound using the Wasserstein metric. Let $(\mathcal{X}, \mathcal{F}, \rho)$ be a general measurable metric space, the distance of a stochastic process $\{X(t)\}$ on $(\mathcal{X}, \mathcal{F})$ to its stationary distribution π is defined by the KR distance

$$\|\mathcal{L}_x(X(t)) - \pi\|_{\mathrm{KR}} := \sup_{g \in \mathrm{Lip}_1^1} |\mathbb{E}[g(X(t))] - \pi(g)| \tag{3.5}$$

where $\mathcal{L}_x(X(t))$ denotes the law of $X(t)$ conditional on starting at $X(0) = x$, $\pi(g) := \int g(x)\pi(\mathrm{d}x)$ is the expected value of g with respect to π, 'KR' stands for 'Kantorovich–Rubinstein', and Lip_1^1 is the set of all functions g from $\mathcal{X}$ to $\mathbb{R}$ with Lipschitz constant no larger than 1 and with $|g(x)| \leq 1$ for all $x \in \mathcal{X}$, i.e.

$$\mathrm{Lip}_1^1 := \{g : \mathcal{X} \to \mathbb{R}, |g(x) - g(y)| \leq \rho(x, y), \forall x, y \in \mathcal{X}, |g| \leq 1\}. \tag{3.6}$$

Note that the KR distance defined in Eq. (3.5) is exactly the 1-st Wasserstein metric. Then it can be shown that the π-average of the KR distance to stationarity from all initial states $X(0)$ in $\mathcal{X}$ is non-increasing, which leads to the following complexity linking proposition.

Proposition 3.1.3. *[RR16, Theorem 1] Let $X^d = \{X^d(t), t \geq 0\}$ be a stochastic process on $(\mathcal{X}, \mathcal{F}, \rho)$, for each $d \in \mathbb{N}$. Suppose X^d converges weakly in the Skorokhod topology as $d \to \infty$ to a càdlàg process X^∞. Assume these processes all have the same stationary distribution π and that X^∞ converges weakly to π. Then for any $\epsilon > 0$, there are $D < \infty$ and $T < \infty$ such that*

$$\mathbb{E}_{X^d(0) \sim \pi} \|\mathcal{L}_{X^d(0)}(X^d(t)) - \pi\|_{KR} < \epsilon, \quad \forall t \geq T, d \geq D. \tag{3.7}$$

Proposition 3.1.3 allows us to bound the convergence of the sequence of processes uniformly over all sufficiently large d, if the sequence of Markov processes converges weakly to a limiting ergodic process. Combining Proposition 3.1.3 with previously-known MCMC diffusion limit results, Roberts and Rosenthal [RR16] prove that the RWM algorithm in d dimensions takes $\mathcal{O}(d)$ iterations to converge to stationarity. However, in [RR16], the target distribution needs to be product i.i.d. with density satisfies all the assumptions of Proposition 3.1.1. Furthermore, the condition Eq. (3.2) is replaced by a stronger condition

$$\int \left[\frac{f'(x)}{f(x)}\right]^{12} f(x)\mathrm{d}x < \infty, \quad \int \left[\frac{f''(x)}{f(x)}\right]^{6} f(x)\mathrm{d}x < \infty. \tag{3.8}$$

3.2 Main results

In this section, we show our main results on optimal scaling of RWM algorithms on general target distributions. We first consider optimal scaling by maximizing ESJD in Section 3.2.1. We show asymptotic form of the ESJD in Theorem 3.2.10 under very mild conditions on the target distribution. Then we show in Theorem 3.2.13 that if we directly maximize the asymptotic ESJD, we can obtain 0.234 as an upper bound of the asymptotically optimal acceptance rate. Next, we show the acceptance rate 0.234 is asymptotically optimal under one more weak law of large number (WLLN) condition on the target distribution in Theorem 3.2.14. In order to give the reader a brief idea that to what extend the class of target distributions can be enlarged. We first present an example of a non-product non-i.i.d. class of distributions, which is a straightforward corollary of our main result in Theorem 3.2.14. Note that our main result includes much more general class of distributions that this simple example. Recall that a (probabilistic) graphical model is a family of probability distributions defined in terms of a directed or undirected graph [Jor04]. Suppose that the statistical model can be represented as a graphical model, then we have the following corollary.

Corollary 3.2.1. *(A Simple Corollary of Theorem 3.2.14) If the following three conditions hold,* 0.234 *is indeed the asymptotic acceptance rate: (i) in the graph representation, each node of the graph has at most* $o(d^{1/4})$ *links; (ii) the target density* π^d *is bounded and* $\log \pi^d$ *has up to the third bounded partial derivatives; (iii) for* $X^d \sim \pi^d$, $\frac{1}{d}\sum_{i=1}^{d}\left(\frac{\partial}{\partial x_i}\log \pi^d(X^d)\right)^2$ *converges to a positive constant as* $d \to \infty$.

In Section 3.2.2, we consider optimal scaling via diffusion limits. We prove the new conditions for weak convergence to diffusion limits in Theorem 3.2.19. We then strengthen this result to consider fixed starting state in Theorem 3.2.21. Finally, in Section 3.2.3, we apply our new result on diffusion limits with fixed starting state to obtain complexity bounds for the RMW algorithm, which is given in Corollary 3.2.23.

Before presenting our main results, we first define a sequence of "sets of typical states".

Definition 3.2.2. We call $\{F_d\}$ a sequence of "sets of typical states" if $\pi^d(F_d) \to 1$.

Next, we enlarge $\{F_d\}$ in different ways, which will be used later for the new conditions on the target.

Definition 3.2.3. For a given sequence of "sets of typical states" $\{F_d\}$, we define

$$F_d^{(i)} := \{(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_d) : \exists (x_1, \ldots, x_d) \in F_d, \text{such that } |y - x_i| < \sqrt{\log d/d}\}. \tag{3.9}$$

Furthermore, we define $F_d^+ := \bigcup_{i=1}^{d} F_d^{(i)}$.

Remark 3.2.4. It is clear from the definitions that $F_d^{(i)}$ is to enlarge the i-th coordinate of $x^d \in F_d$ by covering it with an open interval $(x_i - \sqrt{\log d/d}, x_i + \sqrt{\log d/d})$; F_d^+ is the union of $F_d^{(i)}, i = 1, \ldots, d$. Then clearly we have $F_d \subseteq F_d^{(i)} \subseteq F_d^+$.

Finally, we introduce the idea of "neighborhoods" of a coordinate, which is later used to capture the correlation among different coordinates. We use $\mathcal{H}_i$ to denote a collection of coordinates which are called "neighborhoods" of coordinate i. That is, $\mathcal{H}_i \subseteq \{1, \ldots, d\}$. We also assume $i \in \mathcal{H}_i$. Although the definition of the set $\mathcal{H}_i$ is quite arbitrary, we expect that $j \in \mathcal{H}_i$ implies the coordinates i and j are correlated even conditional on all other coordinates. This idea of "neighborhoods" become clearer if the

target distribution comes from a model which can be written as a probabilistic graphical model [Jor04]. For a graphical model, it is convenient to define the "neighborhood" $j \in \mathcal{H}_i$ if there is an edge between nodes i and j. In this definition, clearly $j \notin \mathcal{H}_i$ implies that the two coordinates i and j are conditional independent given all the other $d-2$ coordinates.

3.2.1 Optimal scaling for maximizing ESJD

Suppose $\{F_d\}$ is a sequence of "sets of typical states" and $\{\mathcal{H}_i\}$ are collections of "neighborhoods" for each coordinate. Throughout the chapter, we assume $\sup_{i\in\{1,\dots,d\}} |\mathcal{H}_i| < l_d$ where $l_d = o(d)$.

Remark 3.2.5. For graphical models, if we define $\mathcal{H}_i$ as the collection of nodes that is directly connected to i by an edge, then $l_d = o(d)$ rules out "dense graphs" for which $l_d \propto d$.

Now we introduce the first assumption A1 on the target π^d.

$$\sup_{(i,j):j\notin\mathcal{H}_i} \sup_{x^d\in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(1), \qquad \sup_{(i,j):j\in\mathcal{H}_i} \sup_{x^d\in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d}). \tag{A1}$$

Remark 3.2.6. For graphical models, if node i is not directly connected to node j, we always have $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = 0$. Therefore, in order to make A1 hold, it suffices to check for each edge of the graph, say (i,j), that $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d})$. Since we have assumed $l_d = o(d)$, this is a very weak condition. For example, A1 holds for all graphical models with bounded second partial derivatives.

Next, we denote the conditional density of the i-th and j-th coordinates, given all the other coordinates fixed, by $\pi_{i,j|-i-j} := \pi^d(x_i, x_j \,|x_{-i-j})$ where x_{-i-j} with $i<j$ denotes all coordinates of x^d other than the i-th, and j-th coordinates, i.e.

$$x_{-i-j} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_{j-1}, x_{j+1}, \dots, x_d).$$

Note that $\pi_{i,j|-i-j}$ is a probability measure in $\mathbb{R}^2$. Then we introduce the next assumption A2 on the target as follows.

$$\sup_{(i,j):j\notin\mathcal{H}_i} \sup_{\{x_{-i-j}:x^d\in F_d\}} \int \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_i^2} \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_j^2} \frac{1}{\pi_{i,j|-i-j}} \mathrm{d}x_i \mathrm{d}x_j = o(1). \tag{A2}$$

Remark 3.2.7. The assumption A2 is very weak, since it is only to require that the target has a "flat tail". To see this, consider the target distribution π^d has the special i.i.d. product form of Eq. (3.1), then A2 reduces to

$$\int \frac{\partial^2 f(x_i) f(x_j)}{\partial x_i^2} \frac{\partial^2 f(x_i) f(x_j)}{\partial x_j^2} \frac{1}{f(x_i)f(x_j)} \mathrm{d}x_i \mathrm{d}x_j = \left(\int \frac{\mathrm{d}^2 f(x)}{\mathrm{d}x^2} \mathrm{d}x \right)^2 = 0, \tag{3.10}$$

when f has a "flat tail" so that $\frac{\mathrm{d}f(x)}{\mathrm{d}x} \to 0$ when $|x| \to \infty$. Similarly, for graphical models, if there is no edge between i and j, then when π^d has "flat tail" we have $\int \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_i^2} \frac{\partial^2 \pi_{i,j|-i-j}}{\partial x_j^2} \frac{1}{\pi_{i,j|-i-j}} \mathrm{d}x_i \mathrm{d}x_j = 0$.

The next assumption is about conditions on the third partial derivatives.

$$\begin{aligned} &\sup_{(i,j):j\notin\mathcal{H}_i}\sup_{x^d\in\mathbb{R}^d}\frac{\partial^3\log\pi^d(x^d)}{\partial x_i^2\partial x_j}=o(1), \quad \sup_{(i,j):j\in\mathcal{H}_i}\sup_{x^d\in\mathbb{R}^d}\frac{\partial^3\log\pi^d(x^d)}{\partial x_i^2\partial x_j}=o(d/l_d),\\ &\sup_i\sup_{x^d\in\mathbb{R}^d}\frac{\partial^3\log\pi^d(x^d)}{\partial x_i^3}=o(d^{1/2}), \quad \sum_{i\neq j\neq k}\left(\sup_{x^d\in\mathbb{R}^d}\left|\frac{\partial^3\log\pi^d(x^d)}{\partial x_i\partial x_j\partial x_k}\right|\right)=o(d^{3/2}). \end{aligned} \tag{A3}$$

Remark 3.2.8. We consider graphical models that satisfy A3. The first three equations of A3 are similar to A1 and they hold for all graphical models with bounded third partial derivatives. Recall that, in graph theory, a n-clique of a graph is a fully-connected subset of nodes of the graph with cardinality n. The last equation of A3 then involves the number of 3-cliques in the graph. Note that for many realistic hierarchical models, there are no 3-cliques for the corresponding graphs, which implies $\sum_{i\neq j\neq k}\left|\frac{\partial^3\log\pi^d(x^d)}{\partial x_i\partial x_j\partial x_k}\right|=0$. Even for the worst case, considering a graph that has d nodes and each has l_d neighbors, since there are $dl_d/2$ links, the number of 3-cliques is at most $\binom{l_d}{2}d/3=\mathcal{O}(l_d^2d)$. Therefore, A3 holds for any graphical model with $l_d=o(d^{1/4})$ and bounded third partial derivatives.

The next assumption is the last assumption before our first main result. We first define a quantity which measures the "roughness" of $\log\pi^d$.

$$I_d(x^d):=\frac{1}{d}\sum_{i=1}^d\left(\frac{\partial}{\partial x_i}\log\pi^d(x^d)\right)^2. \tag{3.11}$$

Similarly, we can consider $I_d(X^d)$ where $X^d\sim\pi^d$ as a random variable. Later we will see that it turns out that $I_d(X^d)$ is a key quantity for optimal scaling results. Assumption A4 is as follows.

There exists α with $0<\alpha<1/2$ such that

$$\sup_i\sup_{x^d\in F_d^{(i)}}\frac{\partial\log\pi^d(x^d)}{\partial x_i}=\mathcal{O}(d^\alpha), \quad \sup_{x^d\in F_d^+}\pi^d(x^d)=o(d^{1/2-\alpha}), \quad \sup_{x^d\in F_d^+}1/I_d(x^d)=\mathcal{O}(d^{\alpha/2}). \tag{A4}$$

Remark 3.2.9. For A4, the first two conditions do not even require π^d and the first partial derivative of $\log\pi^d$ to be bounded. Thus, they are quite weak. For the last condition, although the mode of π^d is ruled out from F_d^+, the condition can hold as long as $\sup_i\sup_{x^d\in F_d^{(i)}}\frac{\partial\log\pi^d(x^d)}{\partial x_i}=\mathcal{O}(d^{\alpha/2})$ and $I_d(X^d)$ is tight. That is, $\forall 0<\epsilon<1$, there exists $K_\epsilon>0$ such that $\mathbb{P}(I_d(X^d)>K_\epsilon)<1-\epsilon)$. To see this, one can choose F_d using the tightness such that $\sup_{x^d\in F_d}1/I_d(x^d)=\mathcal{O}(d^{\alpha/2})$. Then we can replace F_d by F_d^+ since $\inf_{x^d\in F_d}I_d(x^d)-\inf_{x^d\in F_d^+}I_d(x^d)=\mathcal{O}(d^{\alpha/2}(\log d)^{1/2}d^{-1/2})=o(d^{-1/4})=o(d^{-\alpha/2})$. Note that $I_d(X^d)$ being tight is a very reasonable assumption, since if $I_d(X^d)$ is not tight, the target π^d becomes "flat" at almost every state x^d.

We are now ready to present our first main result using the assumptions A1, A2, A3, and A4. We establish the following results on asymptotic ESJD and asymptotic acceptance rate.

Theorem 3.2.10. *(Asymptotic ESJD and acceptance rate) Suppose π^d satisfies A1, A2, A3, and A4,*

then as $d \to \infty$, we have

$$\left| \mathrm{ESJD}(d) - 2\frac{d\ell^2}{d-1}\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]\right| \to 0, \tag{3.12}$$

$$\left| \mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y^d}\left(1\wedge\frac{\pi^d(Y^d)}{\pi^d(X^d)}\right) - 2\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]\right| \to 0, \tag{3.13}$$

where the expectation over Y^d is taken for $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1}I)$ for given $X^d = x^d$.

Proof. See Section 3.4.1. □

Since the assumptions required by Theorem 3.2.10 are very mild, the result of Theorem 3.2.10 holds for a large class of realistic MCMC models. As an example, we give a class of graphical models that all conditions A1, A2, A3, and A4 hold. Therefore, the asymptotic ESJD and acceptance rate by Theorem 3.2.10 hold for this class of graphical models. We will further discuss realistic MCMC models later in Section 3.3.1 and Section 3.3.2.

We give a simple criterion that the assumptions A1, A2, A3, and A4 hold. More discussions and examples are delayed to Section 3.3.

Corollary 3.2.11. *If a graphical model satisfies (i)* either *each node has at most $l_d = o(d^{1/4})$ links* or *the number of 3-cliques of the graph is $o(d^{3/2})$; (ii) $I_d(X^d)$ is tight; (iii) π^d has bounded density and $\log \pi^d$ has up to the third bounded partial derivatives, then the assumptions A1, A2, A3, and A4 hold. Therefore, the asymptotic ESJD and acceptance rate results by Theorem 3.2.10 hold.*

Proof. First, the assumption A1 holds when second partial derivatives of $\log \pi^d$ are bounded. Next, the assumption A2 automatically holds for graphical models. Furthermore, $l_d = o(d^{1/4})$ implies that the number of 3-cliques is $o(d^{3/2})$. Then one can easily verify that the assumption A3 holds using the fact that the third partial derivatives of $\log \pi^d$ are bounded. Finally, the assumption A4 holds since $I_d(X^d)$ is tight. □

Note that Theorem 3.2.10 suggests that under mild conditions on the target distribution, the expected acceptance rate

$$\mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y^d}\left(1\wedge\frac{\pi^d(Y^d)}{\pi^d(X^d)}\right) \to 2\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]. \tag{3.14}$$

Therefore, we can define asymptotic acceptance rate as a function of ℓ as follows.

Definition 3.2.12. (Asymptotic acceptance rate) The asymptotic acceptance rate function is defined by

$$a(\ell) := 2\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]. \tag{3.15}$$

The next theorem shows that if the target distribution satisfies A1, A2, A3 and A4, then if we maximize the asymptotic ESJD, the resulting asymptotic acceptance rate is no larger than 0.234.

Theorem 3.2.13. *Defining the optimal parameter for maximizing the asymptotic ESJD by $\hat{\ell}$, i.e.*

$$\hat{\ell} := \arg\max_{\ell} h(\ell), \quad h(\ell) := 2\ell^2 \mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right], \tag{3.16}$$

then we have $a(\hat{\ell}) \leq 0.234$ (to three decimal places).

Proof. We follow the arguments in [Taw17, Lemma 5.1.4]. First, it can be verified by taking the second derivatives of $h(\ell)$ with respect to ℓ that the maximum of $h(\ell)$ is achieved at ℓ such that $\frac{\partial h(\ell)}{\partial \ell} = 0$. Therefore, the optimal $\hat{\ell}$ satisfies

$$2\mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right] = \mathbb{E}_{X^d \sim \pi^d}\left[\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\Phi'\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right]. \tag{3.17}$$

Therefore, the asymptotic acceptance rate

$$a(\hat{\ell}) = \mathbb{E}_{X^d \sim \pi^d}\left[\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\Phi'\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right] = \mathbb{E}_{X^d \sim \pi^d}\left[-\Phi^{-1}(V)\Phi'\left(\Phi^{-1}(V)\right)\right], \tag{3.18}$$

where $V := \Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)$. By [She06], the function $-\Phi^{-1}(x)\Phi'\left(\Phi^{-1}(x)\right)$ is a concave function for any $x \in (0,1)$. Therefore, we have

$$a(\hat{\ell}) = \mathbb{E}_{X^d \sim \pi^d}\left[-\Phi^{-1}(V)\Phi'\left(\Phi^{-1}(V)\right)\right] \leq -\Phi^{-1}[\mathbb{E}_{X^d \sim \pi^d}(V)]\Phi'\left[\Phi^{-1}(\mathbb{E}_{X^d \sim \pi^d}(V))\right]. \tag{3.19}$$

Defining $m := -\Phi^{-1}[\mathbb{E}_{X^d \sim \pi^d}(V)]$, we can then write $a(\hat{\ell}) = 2\Phi(-m) \leq m\Phi'(-m)$. Finally, it suffices to show that $2\Phi(-m) \leq m\Phi'(-m)$ implies $2\Phi(-m) \leq 0.234$ (to three decimal places). Note that the function $x^2\Phi(-x)$ is maximized at $\hat{m}$ such that $2\Phi(-\hat{m}) = \hat{m}\Phi'(-\hat{m}) \approx 0.234$. By [Taw17, Lemma 5.1.4], the function $2\Phi(-x) - x\Phi'(-x)$ is positive for $x < \hat{m}$ and negative for $x > \hat{m}$. Therefore, $2\Phi(-m) \leq m\Phi'(-m)$ implies that $m > \hat{m}$. Since $\Phi(-x)$ is monotonically decreasing with x, we have $a(\hat{\ell}) = 2\Phi(-m) \leq 2\Phi(-\hat{m}) \approx 0.234$. □

The next result is our main result for optimal scaling by maximizing ESJD. Defining the following WLLN condition for the target π^d:

$$I_d(X^d) - \bar{I}_d \to 0 \quad \text{in probability} \tag{A5}$$

where $X^d \sim \pi^d$ and $\bar{I}_d := \mathbb{E}_{X^d \sim \pi^d}[I_d(X^d)]$, we show that if the target distribution π^d satisfies A1, A2, A3, A4, and the WLLN assumption in A5, then the acceptance rate 0.234 is asymptotically optimal.

Theorem 3.2.14. *(Optimal scaling for maximizing ESJD) Suppose the target distribution π^d satisfies A1, A2, A3, A4, and A5. Then the asymptotic optimal acceptance rate $a(\hat{\ell}) \approx 0.234$ (to three decimal places).*

Proof. By convexity of the function $\Phi(-x)$ when $x \geq 0$, we can immediately obtain a lower bound

$$\ell^2 \mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right] \geq \ell^2\left[\Phi\left(-\frac{\ell\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}{2}\right)\right]. \tag{3.20}$$

Under A5, this lower bound is asymptotically tight. Therefore, as $d \to \infty$, according to [RGG97], we have (to two decimal places)

$$\hat{\ell} \to \frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}, \quad h(\hat{\ell}) \to \frac{1.3}{\left(\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]\right)^2}. \tag{3.21}$$

The acceptance rate which maximizing the asymptotic ESJD is

$$a(\hat{\ell}) = 2\mathbb{E}_{X^d \sim \pi^d}\left[\Phi\left(-\frac{\hat{\ell}\sqrt{I_d(X^d)}}{2}\right)\right] \to 2\Phi\left(-\frac{\hat{\ell}\mathbb{E}_{X^d \sim \pi^d}\sqrt{I_d(X^d)}}{2}\right) \tag{3.22}$$

$$\approx 2\Phi\left(-\frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}\frac{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I_d(X^d)}]}{2}\right) = 2\Phi(-1.19) \approx 0.234. \tag{3.23}$$

□

Remark 3.2.15. Comparing the results of Theorem 3.2.13 and Theorem 3.2.14, it is clear that the "roughness" of π^d, $I_d(X^d)$, is the key quantity which determines the optimal acceptance rate $a(\hat{\ell}) \leq 0.234$ when only the tightness of $I_d(X^d)$ can be verified, or $a(\hat{\ell}) \approx 0.234$ when the concentration of $I_d(X^d)$ as defined in A5 can be verified. We will later demonstrate how to verify A5 for some realistic MCMC models in Section 3.3.1 and Section 3.3.2.

3.2.2 Optimal scaling via diffusion limits

In this subsection, we consider sufficient conditions on π^d for establishing weak convergence of diffusion limits. As we discussed before, establishing such results gives stronger guarantee for optimal scaling than maximizing ESJD. However, it also requires stronger conditions on the target distribution. As we will see in the following, we need to strengthen assumptions A2, A3, A4, A5 and add one more assumption A6.

We first strengthen A2 to a new assumption A2+ as follows.

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\int\left(\frac{\partial^2\pi^d}{\partial x_i^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2\pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2\pi^d}{\partial x_k^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d = \mathcal{O}(d^{2-\delta}) \tag{A2+}$$

for some $\delta > 0$.

Remark 3.2.16. The new assumption A2+ is stronger than A2 but is still very mild. To see this, we consider graphical models as examples. For graphical models with d nodes each with $\mathcal{O}(l_d)$ links, there are at most $\mathcal{O}(dl_d^2)$ 3-cliques. Therefore, A2+ holds for any graphical model with $l_d = o(d^{1/2-\delta})$ and bounded second partial derivatives of $\log \pi^d$. Note that this is only for the worst case, as many realistic graphical models do not have 3-cliques.

Next, we slightly strengthen A3 and A4 to A3+ and A4+.

$$\begin{aligned}&\sup_{(i,j):j\notin\mathcal{H}_i}\sup_{x^d\in\mathbb{R}^d}\frac{\partial^3\log\pi^d(x^d)}{\partial x_i^2\partial x_j} = o(1), \quad \sup_{(i,j):j\in\mathcal{H}_i}\sup_{x^d\in\mathbb{R}^d}\frac{\partial^3\log\pi^d(x^d)}{\partial x_i^2\partial x_j} = o(\sqrt{d/l_d}),\\ &\sum_{i\neq j\neq k}\left(\sup_{x^d\in\mathbb{R}^d}\left|\frac{\partial^3\log\pi^d(x^d)}{\partial x_i\partial x_j\partial x_k}\right|\right) = o(d^{3/2}).\end{aligned} \tag{A3+}$$

Suppose exists $0 < \alpha < 1/2$ that

$$\begin{aligned} &\sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i^2} = o(d^\alpha), \quad \sup_i \sup_{x^d \in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^{\alpha/2}), \\ &\sup_{x^d \in F_d^+} \pi^d(x^d) = o(d^{1/2-\alpha}), \quad \sup_{x^d \in F_d^+} 1/I_d(x^d) = \mathcal{O}(d^{\alpha/4}). \end{aligned} \tag{A4+}$$

Furthermore, we strengthen the WLLN condition A5 to the following A5+.

$$\sup_{x^d \in F_d^+} \left| I_d(x^d) - \bar{I} \right| \to 0 \tag{A5+}$$

where $\bar{I} := \lim_{d\to\infty} \bar{I}_d$ exists.

Remark 3.2.17. A3+ is only slightly stronger than A3 on the rates. A4+ also includes a new condition on the rate of $\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i^2}$ which is quite weak. A5+ requires any sequence $(x^1, x^2, \dots, x^d, \dots)$ where $x^i \in F_i^+$ converges to the same limit $\bar{I}$, so it is (slightly) stronger than WLLN condition in A5. It will become clear in the proof of Theorem 3.2.19 that A5+ is to ensure the speed measure of the diffusion process $h(\ell)$ does not depend on the state x^d.

Finally, we define a new assumption A6 on the target distribution. Roughly speaking, the new assumption is to require the first coordinate of π^d is asymptotically independent with the rest.

$$\lim_{d\to\infty} \sup_{x^d \in F_d^+} \left| \frac{\mathrm{d}}{\mathrm{d}x_1} \left[\log \pi^d(x_1 \,|\, x_{-1}) - \log \tilde{\pi}(x_1) \right] \right| = 0, \tag{A6}$$

where $x_{-1} := (x_2, \dots, x_d)$, $\tilde{\pi}$ is a one-dimensional density and $(\log \tilde{\pi})'$ is Lipschitz continous.

Remark 3.2.18. Note that A6 is a strong condition, which may not be satisfied for many realistic MCMC models. However, it might be necessary in order to get a one-dimensional diffusion limit for the first coordinate. In the proof of the optimal scaling via diffusion limits result in Theorem 3.2.19, the assumption A6 is to ensure the SDE for the first coordinate x_1 doesn't depend on the values of other coordinates. Furthermore, although we do not pursue in this work, if in A6 we instead assume not just the first component but a finite collection of components are asymptotically independent from the rest, a version of weak convergence to multi-dimensional diffusion limits could be obtained following similar arguments as the proof of the one-dimensional diffusion limit case in Theorem 3.2.19.

Now we are ready for the main result of optimal scaling via diffusion limits, which is given in Theorem 3.2.19. Comparing with the assumptions in Theorem 3.2.14, the new sufficient conditions for diffusion limits include strengthening A2 to A2+, A3 and A4 to A3+ and A4+, A5 to A5+, and adding A6. We also require slightly stronger condition on the sequence of "sets of typical states" $\{F_d\}$.

Theorem 3.2.19. *(Optimal scaling via diffusion limits) Suppose the sequence $\{F_d\}$ satisfies $\pi^d(F_d^c) = \mathcal{O}(d^{-1-\delta})$ for some $\delta > 0$, the target distribution π^d satisfies A1, A2+, A3+, A4+, A5+, and A6, then for $U^d(t) := X_1^d(\lfloor dt \rfloor)$, as $d \to \infty$, we have $U^d \Rightarrow U$, where $\Rightarrow$ denotes weak convergence in Skorokhod topology, and U satisfies the Langevin SDE*

$$\mathrm{d}U(t) = (h(\ell))^{1/2} \mathrm{d}B(t) + h(\ell) \frac{\tilde{\pi}'(U(t))}{2\tilde{\pi}(U(t))} \mathrm{d}t, \tag{3.24}$$

where $h(\ell) := 2\ell^2\Phi(-\ell\sqrt{\bar{I}}/2)$ *is the speed measure for the diffusion process.*

Proof. See Section 3.4.3. □

Remark 3.2.20. Note that Theorem 3.2.19 allows dependent coordinates on the target distribution, which is much more general than the product i.i.d. condition. The only strong assumption is A6 which requires the first coordinate is asymptotically independent with other coordinates.

Next, we present another result with slightly stronger conditions, which allows the RWM algorithm to start at a fixed state. This stronger convergence result later allows us to establish a complexity bound for the RMW algorithm in Section 3.2.3 Let $X^d = \{X^d(t), t \geq 0\}$ for $d \in \mathbb{N}$ be the RWM processes defined earlier. Without loss of generality, suppose $\{X^d, d = 1, 2, \dots\}$ are defined in a common measurable metric space $(\mathbb{R}^\infty, \mathcal{F}, \rho)$ as independent processes.

Theorem 3.2.21. *(Optimal scaling via diffusion limits with fixed starting state) Suppose* X_1^d *converges weakly in the Skorokhod topology as* $d \to \infty$ *to a càdlàg process* X_1^∞*. Moreover, assume these processes* $\{X^d, d = 1, 2, \dots\}$ *all have the same marginal stationary distribution* π_1 *for the first coordinate and that the first coordinate of* X^∞ *converges weakly to* π_1*. Suppose the sequence* $\{F_d\}$ *satisfies* $\pi^d(F_d^c) = \mathcal{O}(d^{-2-\delta})$ *for some* $\delta > 0$*, the target distribution* π^d *satisfies A1, A3+, A4+, A5+, and A6. We strengthen A2+ to the following condition*

$$\sum_{i,j,k,l,m\in\{2,\dots,d\}} \int \left(\frac{\partial^2\pi_{-1}}{\partial x_i^2}\cdot\frac{\partial^2\pi_{-1}}{\partial x_j^2}\cdot\frac{\partial^2\pi_{-1}}{\partial x_k^2}\cdot\frac{\partial^2\pi_{-1}}{\partial x_l^2}\cdot\frac{\partial^2\pi_{-1}}{\partial x_m^2}\right)\left(\frac{1}{\pi_{-1}}\right)^5 \pi^d \mathrm{d}x^d = \mathcal{O}(d^{3-6\delta}). \quad \text{(A2++)}$$

Then as $d \to \infty$*, we have* ${}_xU^d \Rightarrow {}_xU$*, where* ${}_xU^d(t) := (X_1^d(\lfloor dt \rfloor) \,|\, X_1^d(0) = x)$ *is the first coordinate of the RWM algorithm sped up by a factor of* d*, conditional on starting at the state* x*, and* ${}_xU$ *is the limiting ergodic Langevin diffusion* U *in Eq. (3.24) also conditional on starting at* x*.*

Proof. See Section 3.4.5. □

Remark 3.2.22. The new assumption A2++ is stronger than A2+ but is still not strong. To see this, for graphical models with d nodes, each with $\mathcal{O}(l_d)$ links, we have at most $\mathcal{O}(dl_d^2)$ 3-cliques. Under flat tail assumptions, at most $\mathcal{O}(d^2 l_d^3)$ terms in the summation in A2++ is not zero. Therefore, A2++ holds for any graphical model with $l_d = o(d^{1/3-2\delta})$ and bounded second partial derivatives of $\log \pi^d$. Note that this is only for the worst case, as many realistic graphical models do not have 3-cliques.

3.2.3 Complexity bounds via diffusion limits

In the following, by combing Theorem 3.2.21 and Proposition 3.1.3, we present a complexity bound for the RWM algorithm which holds for much more general target distributions comparing with [RR16]. More specifically, if the target distribution satisfies the conditions given in Theorem 3.2.21 which allows dependent coordinates of the target distribution, the RWM algorithm in d dimensions takes $\mathcal{O}(d)$ iterations to converge to stationarity.

Corollary 3.2.23. *(Complexity bound for RWM algorithms) Under the conditions of Theorem 3.2.21, for any* $\epsilon > 0$*, there exists* $D < \infty$ *and* $T < \infty$*, such that*

$$\mathbb{E}_{X_1^d(0)\sim\pi_1} \|\mathcal{L}_{X_1^d(0)}(X_1^d(\lfloor dt \rfloor)) - \pi_1\|_{KR} < \epsilon, \quad \forall t \geq T, d \geq D, \tag{3.25}$$

where π_1 denotes the marginal stationary distribution of the first coordinate.

Proof. The result directly comes from Proposition 3.1.3 and Theorem 3.2.21. □

3.3 Examples and applications

In this section, we further discuss examples and applications of the main results in Section 3.2. We first discuss in Section 3.3.1 on verifying the assumptions of Theorem 3.2.14 for realistic MCMC models. We have explained in Remarks 3.2.6 to 3.2.9 that A1, A2, A3, and A4 are typically very weak conditions and they hold for some classes of graphical models. However, as discussed in Remark 3.2.15, the assumption A5 may need to be verified case by case. Particularly, in order to satisfy A5, we may need to make additional assumptions on the observed data. Fortunately, we show by a simple Gaussian example in Example 3.3.1 that, in some cases, A5 can be easily verified without any further assumptions. Then, in Section 3.3.2, we extend the simple Gaussian example in Example 3.3.1 to a more realistic MCMC model in Example 3.3.5 and show it satisfies all the assumptions required by Theorem 3.2.14. Thus, the acceptance rate 0.234 is indeed asymptotically optimal for this realistic MCMC model.

3.3.1 Discussions on Theorem 3.2.14

The optimal scaling result for maximizing ESJD in Theorem 3.2.14 requires one to verify that the target distribution satisfies A1, A2, A3, A4, and A5. We discuss how to verify the conditions on the target distribution required by Theorem 3.2.14 in practice. We explain that A1, A2, A3 and A4 are quite mild and usually easy to be verified. Therefore, we usually only need to focus on the WLLN condition in A5, which might be difficult to check in practice. Throughout this subsection, we demonstrate verification of all the assumptions by a simple Gaussian example, which can be seen as a simplified version of typical Bayesian hierarchical models.

Example 3.3.1. (A Gaussian example) Consider a simple Gaussian MCMC model

$$\begin{aligned} Y_{ij} \mid \theta_{ij} &\sim \mathcal{N}(\theta_{ij}, 1), \quad i, j \in \{1, \dots, n\} \\ \theta_{ij} \mid \mu_j &\sim \mathcal{N}(\mu_j, 1), \quad i \in \{1, \dots, n\} \\ \mu_j \mid \nu &\sim \mathcal{N}(\nu, 1) \\ \nu &\sim \text{ flat prior on } \mathbb{R}, \end{aligned} \tag{3.26}$$

where $\{Y_{ij}\}_{i,j=1}^n$ are the observed data, and $x^d = (\nu, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$ are parameters. Note that we have the number of parameters $d = n^2 + n + 1$ in this example. The target distribution (i.e. the posterior distribution) satisfies

$$\pi^d(x^d) = \mathbb{P}(x^d \mid \{Y_{ij}\}_{i,j=1}^n) \propto \prod_{j=1}^n \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu_j - \nu)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\theta_{ij} - \mu_j)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_{ij} - \theta_{ij})^2}{2}}. \tag{3.27}$$

Note that the hyperparameters ν is conditionally independent given $\{\theta_{ij}\}$. Therefore, ν is only directly dependent with n coordinates $\{\mu_j\}_{j=1}^n$. We can define the "neighborhoods" of ν using the collection of $\mu_j, j = 1, \dots, n$. Similarly, μ_j is directly dependent with ν and $\{\theta_{ij}\}_{i=1}^n$ and θ_{ij} is directly dependent

with μ_j. Therefore, if we choose the directly dependent coordinates as "neighborhoods", we have $l_d = n + 1 = \mathcal{O}(d^{1/2})$.

Verifying A1 to A4

First of all, the two conditions for $(i, j) : j \neq \mathcal{H}_i$ in A1 and A3 hold trivially for graphical models. Furthermore, in Example 3.3.1, the parameter ν is conditional independent with all θ_{ij} and the corresponding conditional posterior distributions all have Gaussian tails, which implies A2 holds for any pair of coordinates (ν, θ_{ij}). Similarly, one can easily verify the assumption holds for other pairs of parameters.

Next, all the conditions on the third partial derivatives of $\log \pi^d$ hold, since there is no 3-cliques. Moreover, in Example 3.3.1, we have $l_d = \mathcal{O}(d^{1/2})$. The second partial derivative is $\mathcal{O}(1)$, and the density π^d is bounded, so the following conditions hold without the need of choosing $\{F_d\}$:

$$\sup_{(i,j):j\in\mathcal{H}_i} \sup_{x^d\in F_d^+} \frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} = o(\sqrt{d/l_d}), \qquad \sup_{x^d\in F_d^+} \pi^d(x^d) = o(d^{1/2-\alpha}). \tag{3.28}$$

Finally, the last two conditions are almost immediately true once A5 has been verified:

$$\sup_{i\in\{1,\dots,d\}} \sup_{x^d\in F_d^+} \frac{\partial \log \pi^d(x^d)}{\partial x_i} = \mathcal{O}(d^{\alpha}), \qquad \sup_{x^d\in F_d^+} 1/I(x^d) = \mathcal{O}(d^{\alpha/2}). \tag{3.29}$$

To see this, under A5, we have $\frac{1}{d}\sum_{i=1}^d \left(\frac{\partial}{\partial x_i} \log \pi^d(x^d)\right)^2 \to \bar{I}_d$. If $\bar{I}_d \to \bar{I}$ and $\bar{I} > 0$, then we can select constant $K_2 > 0$ small enough such that $\bar{I} > K_2 d^{-\alpha/2} > 0$ then $\bar{I}_d > K_2 d^{-\alpha/2}$ for all large enough d. Next, by choosing the typical set F_d such that for any $x^d \in F_d^+$, we have $\frac{\partial \log \pi^d(x^d)}{\partial x_i} \leq K_1 d^{\alpha}, \quad I_d(x^d) \geq K_2 d^{-\alpha/2}$, where K_1 is a large enough constant. Then it suffices to check if $\{F_d\}$ is a valid sequence of typical sets such that $\pi^d(F_d) \to 1$. For Example 3.3.1, we have $X^d = (\nu, \{\nu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$. We will show later that A5 holds such that under $X^d \sim \pi^d$ we have $\frac{1}{d}\sum_{i=1}^d \left(\frac{\partial}{\partial x_i} \log \pi^d(X^d)\right)^2 \to 3$. For example, we can choose $K_2 = 0.01$, $K_1 = 100$, and the typical set F_d such that, for any $X^d = x^d \in F_d^+$, we have

$$I_d(x^d) > 0.01 n^{-\alpha}, \quad \frac{\partial \log \pi^d}{\partial \nu} = n(\bar{\mu} - \nu) \leq 100 n^{2\alpha}, \tag{3.30}$$

$$\frac{\partial \log \pi^d}{\partial \mu_j} = (n+1)\left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right) \leq 100 n^{2\alpha}, \tag{3.31}$$

$$\frac{\partial \log \pi^d}{\partial \theta_{ij}} = 2\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right) \leq 100 n^{2\alpha}, \tag{3.32}$$

where $\alpha < 1/2$ can be arbitrarily close to $1/2$. Observing that, under $X^d \sim \pi^d$, we have the following conditional distributions.

$$\begin{aligned} \theta_{ij} \mid Y_{ij}, \mu_j &\sim^{\text{indep.}} \mathcal{N}\left(\frac{\mu_j + Y_{ij}}{2}, \frac{1}{2}\right), \quad i, j \in \{1, \dots, n\}, \\ \mu_j \mid \sum_i \theta_{ij}, \nu &\sim^{\text{indep.}} \mathcal{N}\left(\frac{\sum_i \theta_{ij} + \nu}{n+1}, \frac{1}{n+1}\right), \quad i \in \{1, \dots, n\}, \\ \nu \mid \bar{\mu} &\sim \mathcal{N}\left(\bar{\mu}, \frac{1}{n}\right). \end{aligned} \tag{3.33}$$

Then it can be easily verified that $\pi^d(F_d) \to 1$.

Verifying A5

One assumption of Theorem 3.2.14 that could be difficult to verify in practice is A5. It requires the sequence of random variables $\{I_d(X^d)\}$ converge to a sequence of constants in probability. We feel this assumption has to be checked case by case and it is hard to get general sufficient condition for it to hold. For realistic MCMC models, this may require assumptions on the observed data so that the posterior distribution has certain "concentration" properties as $d \to \infty$.

Fortunately, for Example 3.3.1, we can verify that A5 holds without any further assumption on the observed data $\{Y_{ij}\}$. Note that in Example 3.3.1, we have

$$\left(\frac{\partial \log \pi^d}{\partial \nu}\right)^2 = \left(\sum_j (\mu_j - \nu)\right)^2 = n^2 (\bar{\mu} - \nu)^2, \tag{3.34}$$

$$\left(\frac{\partial \log \pi^d}{\partial \mu_j}\right)^2 = \left(\sum_i (\theta_{ij} - \mu_j) - (\mu_j - \nu)\right)^2 = (n+1)^2 \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2, \tag{3.35}$$

$$\left(\frac{\partial \log \pi^d}{\partial \theta_{ij}}\right)^2 = ((Y_{ij} - \theta_{ij}) - (\theta_{ij} - \mu_j))^2 = 4\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2. \tag{3.36}$$

Hence, if suffices to show that, under $X^d = (\nu, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n) \sim \pi^d$, the following three terms converges to some constants in probability or in distribution:

$$\frac{1}{d}\left(\frac{\partial \log \pi^d}{\partial \nu}\right)^2 = \frac{n^2}{n^2+n+1}(\bar{\mu} - \nu)^2, \tag{3.37}$$

$$\frac{1}{d}\sum_j \left(\frac{\partial \log \pi^d}{\partial \mu_j}\right)^2 = \frac{(n+1)^2}{n^2+n+1}\sum_j \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2, \tag{3.38}$$

$$\frac{1}{d}\sum_{ij} \left(\frac{\partial \log \pi^d}{\partial \theta_{ij}}\right)^2 = \frac{4}{d}\sum_{ij}\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2. \tag{3.39}$$

We have observed that the target distribution π^d has conditional independence structure in Eq. (3.33), which immediately leads to

$$(\bar{\mu} - \nu)^2 \to^{\mathbb{P}} 0, \quad \sum_j \left(\frac{\sum_i \theta_{ij} + \nu}{n+1} - \mu_j\right)^2 \to^{\mathbb{P}} 1, \quad \frac{1}{d}\sum_{ij}\left(\frac{Y_{ij} + \mu_j}{2} - \theta_{ij}\right)^2 \to^{\mathbb{P}} \frac{1}{2}. \tag{3.40}$$

Therefore, A5 is satisfied.

Overall, we have checked all the assumptions of Theorem 3.2.14 for our simple Gaussian example. Therefore, by Theorem 3.2.14, we have the following optimal scaling result for Example 3.3.1.

Proposition 3.3.2. *The optimal scaling for Example 3.3.1 by maximizing ESJD is to choose (to two decimal places)* $\hat{\ell} \approx \frac{2.38}{\mathbb{E}_{X^d \sim \pi^d}[\sqrt{I(X^d)}]} \to \frac{2.38}{\sqrt{3}} \approx 1.37$ *and the corresponding asymptotic acceptance rate is (to three decimal places)* 0.234.

3.3.2 Optimal scaling of a realistic MCMC model

We first discuss sufficient conditions for two more classes of graphical models. In Proposition 3.3.3, we give sufficient conditions for the first equation of A1, A2, and the first equation of A3 to hold for one

particular class of graphical models. In Proposition 3.3.4, we give sufficient conditions for A5 to hold for one specific class of graphical models.

First, we consider the class of graphical models represented by the factor graphs:

$$\pi^d(x^d) \propto \prod_{k=1}^{K_d} \psi_k(\{x_i : i \in C_k\}), \tag{3.41}$$

where C_k are cliques, ψ_k are potentials, K_d denotes the number of potentials.

Proposition 3.3.3. *For the class of graphical models represented by Eq. (3.41). Let m_d denotes the maximum number of cliques a coordinate can belong to. If all the potentials ψ_k have "flat tails" in the sense that for all k we have $\frac{\partial \psi_k}{\partial x_i} \to 0$ as $|x_i| \to \infty$ for all $i \in C_k$, and the cardinality of C_k satisfies $\sup_k |C_k| = o(d/m_d)$, then the first equation in A1, A2, and the first equation in A3 hold.*

Next, we consider Bayesian hierarchical modeling where K denotes the number of "layers" or "stages" of the model. We use $\theta^{(k)}, k = 1, \ldots, K$ to denote the parameter vector with length n_k for the k-th layer, where $\theta^{(k)} := (\theta_1^{(k)}, \ldots, \theta_{n_k}^{(k)})$. We consider the special structure of the graphical model such that $\theta^{(k)}$ is only connected to $\theta^{(k-1)}$ and $\theta^{(k+1)}$. Using factor graphs, let $x^d = (\theta^{(1)}, \ldots, \theta^{(K)})$ we can represent the target distribution as

$$\pi^d(x^d) \propto \prod_{k=1}^{K} \psi_k(\theta^{(k-1)}, \theta^{(k)}), \tag{3.42}$$

where $d = \sum_{k=1}^K n_k$, $\{\psi_k\}$ are the potentials, and without loss of generality we assumed $\theta^{(0)}$ to be the observed data.

In the following, we show that A5 hold for the class of graphical models represented by Eq. (3.42) under certain conditions.

Proposition 3.3.4. *For the class of graphical models represented by Eq. (3.42), if $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{n_k}^{(k)})$ are independent conditional on $\theta^{(k-1)}$ and $\theta^{(k+1)}$ and this holds for all k. Moreover, if under $X^d = (\theta^{(1)}, \ldots, \theta^{(K)}) \sim \pi^d$ all the potentials ψ_k satisfy*

$$\sup_{i \in \{1,\ldots,n_k\}} \left| \frac{\partial \log \psi_k}{\partial \theta_i^{(k)}} \right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{d/n_k}\right), \quad \sup_{j \in \{1,\ldots,n_{k-1}\}} \left| \frac{\partial \log \psi_k}{\partial \theta_j^{(k-1)}} \right| = \mathcal{O}_{\mathbb{P}}\left(\sqrt{d/n_{k-1}}\right) \tag{3.43}$$

then A5 holds.

Next, we extend the simple Gaussian example in Example 3.3.1 to a more realistic MCMC model which belongs to both classes of graphical models in Eqs. (3.41) and (3.42) and show that all the assumptions for the optimal scaling result in Theorem 3.2.14 hold.

Example 3.3.5. (A realistic MCMC model) Consider a realistic MCMC model

$$\begin{aligned} Y_{ij} \mid \theta_{ij} &\sim \mathcal{N}(\theta_{ij}, W), \quad i,j \in \{1,\dots,n\} \\ \theta_{ij} \mid \mu_j &\sim \mathcal{N}(\mu_j, V), \quad i \in \{1,\dots,n\} \\ \mu_j \mid \nu &\sim \mathcal{N}(\nu, A) \\ \nu &\sim \text{ flat prior on } \mathbb{R}, \\ A &\sim \mathbf{IG}(a,b), \end{aligned} \tag{3.44}$$

where $x^d = (\nu, A, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n)$ are parameters, $\{Y_{ij}\}$ are the observed data, and a, b, W, V are known constants.

We further assume that the observed data $\{Y_{ij}\}$ is not abnormal so that the posterior of the hyperparameter A concentrates to some unknown constant.

Assumption. *The posterior of the hyperparameter A in Example 3.3.5 concentrates to some unknown constant $A_0 > 0$ as $n \to \infty$.*

Note that this is a very reasonable assumption which implies the MCMC model is not seriously misspecified. We do not discuss sufficient conditions on the observed data $\{Y_{ij}\}_{i,j=1}^n$ for concentration of posterior distribution of A here since it is not the focus of this work. Next, we show that, under this assumption, the realistic MCMC model satisfies all the conditions required for optimal scaling in Theorem 3.2.14. Therefore, the acceptance rate 0.234 is indeed asymptotically optimal for this MCMC model in the sense of maximizing ESJD.

Proposition 3.3.6. *Under the above assumption, the optimal asymptotic acceptance rate for the realistic MCMC model in Example 3.3.5 is (to three decimal places)* 0.234.

Proof. See Section 3.4.6. □

3.4 Appendices

3.4.1 Proof of Theorem 3.2.10

Throughout the proof, for simplicity, we assume the coordinates are linear ordered. The "neighborhoods" of a coordinate is defined by $\mathcal{H}_i := \{j : |i-j| < l_d\}$. Therefore $\sup_{(i,j):j\in\mathcal{H}_i}$ can be simplified to $\sup_{|i-j|<l_d}$ and $\sup_{(i,j):j\notin\mathcal{H}_i}$ can be simplified to $\sup_{|i-j|\geq l_d}$. Note that the use of linear ordering is only for simplifying notations. It is straightforward to extend the proof to the cases of general ordering.

For Theorem 3.2.10, we only prove

$$\left| \mathrm{ESJD}(d) - 2\frac{d\ell^2}{d-1}\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]\right| \to 0, \tag{3.45}$$

since the proof of

$$\left|\mathbb{E}_{X^d\sim\pi^d}\mathbb{E}_{Y^d}\left(1\wedge\frac{\pi^d(Y^d)}{\pi^d(X^d)}\right) - 2\mathbb{E}_{X^d\sim\pi^d}\left[\Phi\left(-\frac{\ell\sqrt{I_d(X^d)}}{2}\right)\right]\right| \to 0 \tag{3.46}$$

follows similarly.

First, we write ESJD as $\mathrm{ESJD}(d) =: \sum_{i=1}^{d} \mathrm{ESJD}_i(d)$, where

$$\mathrm{ESJD}_i(d) := \mathbb{E}_{X^d \sim \pi^d} \mathbb{E}_{Y^d} \left[(Y_i - X_i)^2 \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right]. \tag{3.47}$$

Then it suffices to show that

$$\sup_{i \in \{1,\dots,d\}} \left| \mathrm{ESJD}_i(d) - \frac{2\ell^2}{d-1} \mathbb{E}_{X^d \sim \pi^d} \left[\Phi\left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right] \right| = o(d^{-1}). \tag{3.48}$$

Writing $\mathrm{ESJD}_i(d) = \mathbb{E}_{X^d \sim \pi^d} \mathbb{E}_{Y_i} \left[(Y_i - X_i)^2 \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right]$, it suffices to show that uniformly over $i \in \{1, \dots, d\}$

$$\mathbb{E}_{X^d \sim \pi^d} \left| \mathbb{E}_{Y_i} \left[(Y_i - X_i)^2 \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) \right] - \frac{2\ell^2}{d-1} \Phi\left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right| \tag{3.49}$$

$$= \mathbb{E}_{X^d \sim \pi^d} \left| \mathbb{E}_{Y_i} \left\{ (Y_i - X_i)^2 \left[\mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(X^d)} \right) - 2\Phi\left(-\frac{\ell \sqrt{I_d(X^d)}}{2} \right) \right] \right\} \right| \tag{3.50}$$

$$= o(d^{-1}). \tag{3.51}$$

It then suffices to show

$$\sup_{x^d \in F_d} \left| \mathbb{E}_{Y_i} \left\{ (Y_i - x_i)^2 \mathbb{1}_{y^d(i) \in F_d^{(i)}} \left[\mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) - 2\Phi\left(-\frac{\ell \sqrt{I_d(x^d)}}{2} \right) \right] \right\} \right| \tag{3.52}$$

$$\leq \mathbb{E}_{Y_i} \left\{ (Y_i - x_i)^2 \sup_{y^d(i) \in F_d^{(i)}, x^d \in F_d} \left| \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) - 2\Phi\left(-\frac{\ell \sqrt{I_d(x^d)}}{2} \right) \right| \right\} = o(d^{-1}), \tag{3.53}$$

where $y^d(i) := (x_1, \dots, x_{i-1}, Y_i, x_{i+1}, \dots, x_d)$. Defining $M_{x^d}^{(i)}(Y_i) := \mathbb{E}_{Y_{-i}} \left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right)$, since

$$\log \frac{\pi^d(Y^d)}{\pi^d(x^d)} = \log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(Y_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \tag{3.54}$$

$$= \left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \right) + \log \frac{\pi_{-i}(Y_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, Y_i)}, \tag{3.55}$$

we can write

$$M_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right] = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp\left(\log \frac{\pi^d(Y^d)}{\pi^d(x^d)} \right) \right] \tag{3.56}$$

$$= \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp\left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} + \log \frac{\pi_{-i}(Y_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, Y_i)} \right) \right]. \tag{3.57}$$

Note that the expectation is taken over Y_{-i} and only the last term, $\log \frac{\pi_{-i}(Y_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, Y_i)}$, involves Y_{-i}.

In the following, we then first focus on approximating $\log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)}$ for given $x^d \in F_d^+$. Since $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1} I)$, we first approximate $\log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)}$ by the first two terms of its Taylor expansion.

Define

$$m_1^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2}(Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i}), \tag{3.58}$$

where

$$(\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) := \sum_{j \in \{1,\dots,d\}, j \neq i} \frac{\partial \log \pi_{-i}(x_{-i} \,|\, x_i)}{\partial x_j} (Y_j - x_j) \tag{3.59}$$

and $[\nabla^2 \log \pi_{-i}]$ denotes the $(d-1) \times (d-1)$ matrix with elements

$$\left\{ \frac{\partial^2 \log \pi_{-i}(x_{-i} \,|\, x_i)}{\partial x_j \partial x_k} \right\}_{j,k \in \{1,\dots,d\}, j \neq i, k \neq i}.$$

Then, we have the following result.

Lemma 3.4.1. *Uniformly over $i \in \{1, \dots, d\}$, we have*

$$\sup_{x^d \in F_d^+} \mathbb{E}_{Y_{-i}} \left[\left| m_1^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \right| \right] \to 0. \tag{3.60}$$

Proof. See Section 3.4.2. □

Next, we approximate the second order term of the Taylor approximation $\frac{1}{2}(Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})$ by a non-random term $\frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}$.

Lemma 3.4.2. *Uniformly over $i \in \{1, \dots, d\}$, we have*

$$\sup_{x^d \in F_d^+} \mathbb{E}_{Y_{-i}} \left[\left| (Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i}) - \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2} \right| \right] \to 0. \tag{3.61}$$

Proof. See Section 3.4.2. □

Defining

$$m_2^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2} \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}, \tag{3.62}$$

we have

$$m_2^{(i)}(Y_{-i}, x^d) \sim \mathcal{N}\left(\ell^2 S_d^{(i)}/2, \ell^2 R_d^{(i)} \right), \tag{3.63}$$

where

$$R_d^{(i)} := \frac{1}{d-1} \sum_{j \neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \,|\, x_i)}{\partial x_j} \right)^2, \quad S_d^{(i)} := \frac{1}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}(x_{-i} \,|\, x_i)}{\partial x_j^2}. \tag{3.64}$$

Next, we show we can approximate $S_d^{(i)}$ by $-R_d^{(i)}$.

Lemma 3.4.3. *There exists a sequence of subsets of states $\{F'_d\}$, such that $\pi^d(F'_d) \to 1$ and*

$$\sup_{i\in\{1,\dots,d\}} \sup_{x^d\in F'_d} \left| R_d^{(i)} + S_d^{(i)} \right| \to 0. \tag{3.65}$$

Proof. See Section 3.4.2. □

Now defining

$$m_3^{(i)}(Y_{-i}, x^d) := (\nabla \log \pi_{-i})^T (Y_{-i} - x_{-i}) + \frac{1}{2}\frac{\ell^2}{d-1} \sum_{j\neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \,|\, x_i)}{\partial x_j} \right)^2, \tag{3.66}$$

we have

$$m_3^{(i)}(Y_{-i}, x^d) \sim \mathcal{N}\left(-\ell^2 R_d^{(i)}/2, \ell^2 R_d^{(i)} \right). \tag{3.67}$$

By triangle inequality, we can write

$$\left| m_3^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \right| \leq \left| m_1^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \right| \tag{3.68}$$

$$+ \left| m_2^{(i)}(Y_{-i}, x^d) - m_1^{(i)}(Y_{-i}, x^d) \right| \tag{3.69}$$

$$+ \left| m_3^{(i)}(Y_{-i}, x^d) - m_2^{(i)}(Y_{-i}, x^d) \right|. \tag{3.70}$$

Therefore, using Lemmas 3.4.1 to 3.4.3, we get

$$\sup_{i\in\{1,\dots,d\}} \sup_{x^d\in F_d^+\cap F'_d} \mathbb{E}_{Y_{-i}} \left[\left| m_3^{(i)}(Y_{-i}, x^d) - \log \frac{\pi_{-i}(Y_{-i} \,|\, x_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} \right| \right] \to 0. \tag{3.71}$$

Next, we abuse the notation a little bit by defining

$$R_d^{(i)}(y) := \frac{1}{d-1} \sum_{j\neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \,|\, x_i = y)}{\partial x_j} \right)^2. \tag{3.72}$$

Then by the definition of $m_3^{(i)}$, we replace x^d by $y^d(i) = (x_1, \dots, x_{i-1}, Y_i, x_{i+1}, \dots, x_d)$, which yields

$$m_3^{(i)}(Y_{-i}, y^d(i)) = (\nabla \log \pi_{-i}(x_{-i} \,|\, Y_i))^T (Y_{-i} - x_{-i}) \tag{3.73}$$

$$+ \frac{1}{2}\frac{\ell^2}{d-1} \sum_{j\neq i} \left(\frac{\partial \log \pi_{-i}(x_{-i} \,|\, Y_i)}{\partial x_j} \right)^2. \tag{3.74}$$

Then, we have

$$m_3^{(i)}(Y_{-i}, y^d(i)) \sim \mathcal{N}\left(-\ell^2 R_d^{(i)}(Y_i)/2, \ell^2 R_d^{(i)}(Y_i) \right). \tag{3.75}$$

Recall that $M_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp\left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} + \log \frac{\pi_{-i}(Y_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, Y_i)} \right) \right]$, defining

$$\hat{M}_{x^d}^{(i)}(Y_i) = \mathbb{E}_{Y_{-i}} \left[1 \wedge \exp\left(\log \frac{\pi_i(Y_i)}{\pi_i(x_i)} + \log \frac{\pi_{-i}(x_{-i} \,|\, Y_i)}{\pi_{-i}(x_{-i} \,|\, x_i)} + m_3^{(i)}(Y_{-i}, y^d(i)) \right) \right], \tag{3.76}$$

we next apply the following two lemmas from [RGG97].

Lemma 3.4.4. *([RGG97, Proposition 2.2]) The function $g(x) = 1 \wedge e^x$ is Lipschitz such that*

$$|g(x) - g(y)| \leq |x - y|, \quad \forall x, y. \tag{3.77}$$

Lemma 3.4.5. *([RGG97, Proposition 2.4]) If $z \sim \mathcal{N}(\mu, \sigma^2)$ then*

$$\mathbb{E}(1 \wedge e^z) = \Phi(\mu/\sigma) + \exp(\mu + \sigma^2/2)\Phi(-\sigma - \mu/\sigma). \tag{3.78}$$

By Lemma 3.4.4 and Eq. (3.71), we have that uniformly over $i \in \{1, \ldots, d\}$

$$\sup_{y^d(i) \in F_d^+ \cap F_d'} \left| M_{x^d}^{(i)}(Y_i) - \hat{M}_{x^d}^{(i)}(Y_i) \right| \to 0. \tag{3.79}$$

Applying Lemma 3.4.5 to $\hat{M}_{x^d}^{(i)}(Y_i)$ yields

$$\hat{M}_{x^d}^{(i)}(Y_i) = \Phi\left(R_d^{(i)}(Y_i)^{-1/2} \left(\ell^{-1} \log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} - \ell R_d^{(i)}(Y_i)/2 \right) \right) \tag{3.80}$$

$$+ \exp\left(\log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} \right) \Phi\left(-\ell R_d^{(i)}(Y_i)^{1/2}/2 - \log \frac{\pi^d(y^d(i))}{\pi^d(x^d)} R_d^{(i)}(Y_i)^{-1/2} \ell^{-1} \right). \tag{3.81}$$

Note that it is easy to check that $\hat{M}_{x^d}^{(i)}(x_i) = 2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2} \right)$. We then show $\hat{M}_{x^d}^{(i)}(x_i)$ converges to $2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2} \right)$.

Lemma 3.4.6.

$$\sup_{i \in \{1, \ldots, d\}} \sup_{x^d \in F_d^+} \left| 2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2} \right) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2} \right) \right| \to 0. \tag{3.82}$$

Proof. See Section 3.4.2. □

Finally, using Taylor expansion together with $\mathbb{E}_{Y_i}(Y_i - x_i)^2 = \ell^2/(d-1)$ and $\mathbb{E}_{Y_i}|Y_i - x_i|^3 = \mathcal{O}(d^{-3/2})$, we have

$$\mathbb{E}_{Y_i}\left\{ (Y_i - x_i)^2 \sup_{y^d(i) \in F_d^+} \left| \hat{M}_{x^d}^{(i)}(Y_i) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2} \right) \right| \right\} \tag{3.83}$$

$$\leq \frac{\ell^2}{d-1} \sup_{x^d \in F_d^+} \left| 2\Phi\left(-\frac{\ell\sqrt{R_d^{(i)}}}{2} \right) - 2\Phi\left(-\frac{\ell\sqrt{I_d(x^d)}}{2} \right) \right| \tag{3.84}$$

$$+ \mathcal{O}(d^{-3/2}) \sup_{y^d(i) \in F_d^+} \left| \frac{\mathrm{d}\hat{M}_{x^d}^{(i)}(y_i)}{\mathrm{d}y_i}(Y_i) \right|. \tag{3.85}$$

For the last term, we have the following lemma.

Lemma 3.4.7.

$$\sup_{i\in\{1,\dots,d\}}\sup_{y^d(i)\in F_d^+}\left|\frac{\mathrm{d}\hat{M}^{(i)}_{x^d}(y_i)}{\mathrm{d}y_i}(Y_i)\right| = o\left(d^{1/2}\right). \tag{3.86}$$

Proof. See Section 3.4.2. □

The proof of Theorem 3.2.10 is completed by applying Lemma 3.4.6 and Lemma 3.4.7.

3.4.2 Proof of lemmas in Section 3.4.1

Proof of Lemma 3.4.1

For $x^d \in F_d^+$, by Taylor expansion and mean value theorem, we have

$$|\log\pi_{-i}(Y_{-i}\,|\,x_i) - \log\pi_{-i}(x_{-i}\,|\,x_i) - m_1(Y_{-i},x^d)| \tag{3.87}$$

$$\le \sup_{\tilde{x}^d\in\mathbb{R}^d}\left|\frac{1}{6}\sum_{j,k,l\neq i}\frac{\partial^3\log\pi^d(\tilde{x}^d)}{\partial x_j\partial x_k\partial x_l}(Y_j-x_j)(Y_k-x_k)(Y_l-x_l)\right|. \tag{3.88}$$

In the above summation, the summation over the cases of $j=k=l$ equals to

$$\sup_{\tilde{x}^d\in\mathbb{R}^d}\left|\frac{\partial^3\log\pi^d(\tilde{x}^d)}{\partial x_j^3}\right|\mathcal{O}(d\mathbb{E}|Y_j-x_j|^3) = o(d^{1/2})\mathcal{O}\left(d(\sqrt{\ell^2/(d-1)})^3\right) = o(1). \tag{3.89}$$

For the cases of $j=k\neq l$, we have

$$\sum_{j=k\neq l}\frac{\partial^3\log\pi_{-i}(\tilde{x}^d)}{\partial x_j^2\partial x_l}(Y_j-x_j)^2(Y_l-x_l) = \sum_j(Y_j-x_j)^2\sum_{l\neq k}\frac{\partial^3\log\pi_{-i}(\tilde{x}^d)}{\partial x_j^2\partial x_l}(Y_l-x_l). \tag{3.90}$$

By Assumption A3, we have $\mathbb{E}\left|\sum_{j\neq l}\frac{\partial^3\log\pi_{-i}(\tilde{x}^d)}{\partial x_j^2\partial x_l}(Y_l-x_l)\right| = \mathcal{O}(l_d/d)o(d/l_d) = o(1)$ since $\frac{\partial^3\log\pi^d(\tilde{x}^d)}{\partial x_j^2\partial x_l}$ goes to zero when $|k-i|>l_d$. Then, by $\mathbb{E}|Y_j-x_j|^2=\mathcal{O}(1/d)$, the summation over all cases of $j=k\neq l$ equals to $d\mathcal{O}_{\mathbb{P}}(1/d)o_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$.

Finally, for $j\neq k\neq l$, it suffices to show

$$\sup_{\tilde{x}^d\in\mathbb{R}^d}\left|\sum_{j\neq k\neq l\neq i}\frac{\partial^3\log\pi_{-i}(\tilde{x}^d)}{\partial x_j\partial x_k\partial x_l}(Y_j-x_j)(Y_k-x_k)(Y_l-x_l)\right| \tag{3.91}$$

$$\le \sum_{i\neq j\neq k\neq l}\left(\sup_{\tilde{x}^d\in\mathbb{R}^d}\left|\frac{\partial^3\log\pi_{-i}(\tilde{x}^d)}{\partial x_j\partial x_k\partial x_l}\right|\right)|(Y_j-x_j)(Y_k-x_k)(Y_l-x_l)| = o_{\mathbb{P}}(1). \tag{3.92}$$

Note that $\{|(Y_j-x_j)(Y_k-x_k)(Y_l-x_l)|\}_{j\neq k\neq l}$ are independent random variables which don't depend on the values of x_j, x_k, x_l, and

$$|(Y_j-x_j)(Y_k-x_k)(Y_l-x_l)| = \mathcal{O}_{\mathbb{P}}\left((\sqrt{\ell^2/(d-1)})^3\right) = \mathcal{O}_{\mathbb{P}}(d^{-3/2}). \tag{3.93}$$

Therefore, the summation for cases $j\neq k\neq l$ is $o_{\mathbb{P}}(1)$ under Assumption A3. We have proven the result for fixed i. Finally, it is easy to check the proof holds uniformly over $i\in\{1,\dots,d\}$.

Proof of Lemma 3.4.2

Lemma 3.4.8. *(Quadratic form of Gaussian random vector) If $z^d \sim \mathcal{N}_d(\mu, \Sigma)$, then*

$$\mathbb{E}(z^T A z) = \operatorname{tr}(A\Sigma) + \mu^T A \mu, \quad \operatorname{var}(z^T A z) = 2\operatorname{tr}(A\Sigma A\Sigma) + 4\mu^T A \Sigma A \mu. \tag{3.94}$$

Note that $Y_{-i} \sim \mathcal{N}_{d-1}(x_{-i}, \frac{\ell^2}{d-1} I)$ and $(Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})$ is a quadratic form of Gaussian random vector. By Lemma 3.4.8,

$$\mathbb{E}\left[(Y_{-i} - x_{-i})^T [\nabla^2 \log \pi_{-i}](Y_{-i} - x_{-i})\right] = \frac{\ell^2}{d-1} \sum_{j \neq i} \frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}. \tag{3.95}$$

Therefore, it suffices to show the variance of the quadratic form goes to zero. Using the assumptions, the variance satisfies

$$\frac{2\ell^4}{(d-1)^2} \operatorname{tr}\left([\nabla^2 \log \pi_{-i}][\nabla^2 \log \pi_{-i}]\right) \tag{3.96}$$

$$= \frac{2\ell^4}{(d-1)^2} \sum_{j \neq i} \sum_{k \neq i} \left(\frac{\partial^2 \log \pi_{-i}}{\partial x_j \partial x_k}\right)^2 \tag{3.97}$$

$$\leq \frac{2\ell^4}{(d-1)^2} \sum_{l=0}^{d-1} \sum_{\{j,k:|j-k|=l\}} \left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{3.98}$$

$$= \frac{2\ell^4}{(d-1)^2} \sum_{l \leq l_d} \sum_{\{j,k:|j-k|=l\}} \left(\frac{\partial^2 \log \pi^d}{\partial x_i \partial x_j}\right)^2 \tag{3.99}$$

$$+ \frac{2\ell^4}{(d-1)^2} \sum_{l > l_d} \sum_{\{j,k:|j-k|=l\}} \left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{3.100}$$

$$\leq \frac{2\ell^4}{(d-1)^2} (d-1) l_d \sup_{|j-k| \leq l_d} \sup_{x^d \in F_d^+} \left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{3.101}$$

$$+ \frac{2\ell^4}{(d-1)^2} (d-1)^2 \sup_{|j-k| > l_d} \sup_{x^d \in F_d^+} \left(\frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k}\right)^2 \tag{3.102}$$

$$= \mathcal{O}(l_d/d) o(d/l_d) + o(1) = o(1), \tag{3.103}$$

where we have used $\sup_{x^d \in F_d^+} \sup_{|j-k| \leq l_d} \frac{\partial^2 \log \pi^d}{\partial x_j \partial x_k} = o(\sqrt{d/l_d})$ from Assumption A1.

Proof of Lemma 3.4.3

Note that

$$R_d^{(i)} + S_d^{(i)} = \frac{1}{d-1}\sum_{j\neq i}\left(\frac{\partial \log \pi_{-i}}{\partial x_j}\right)^2 + \frac{1}{d-1}\sum_{j\neq i}\frac{\partial^2 \log \pi_{-i}}{\partial x_j^2} \tag{3.104}$$

$$= \frac{1}{d-1}\sum_{j\neq i}\left\{\left(\frac{\partial \log \pi^d}{\partial x_j}\right)^2 + \frac{\partial^2 \log \pi^d}{\partial x_j^2}\right\} \tag{3.105}$$

$$= \frac{1}{d-1}\sum_{j\neq i}\left\{\frac{1}{(\pi^d)^2}\left(\frac{\partial \pi^d}{\partial x_j}\right)^2 + \frac{\partial}{\partial x_j}\left(\frac{\partial \log \pi^d}{\partial x_j}\right)\right\} \tag{3.106}$$

$$= \frac{1}{d-1}\sum_{j\neq i}\left\{\frac{1}{(\pi^d)^2}\left(\frac{\partial \pi^d}{\partial x_j}\right)^2 + \frac{\partial}{\partial x_j}\left(\frac{1}{\pi^d}\frac{\partial \pi^d}{\partial x_j}\right)\right\} \tag{3.107}$$

$$= \frac{1}{d-1}\sum_{j\neq i}\left\{\frac{1}{(\pi^d)^2}\left(\frac{\partial \pi^d}{\partial x_j}\right)^2 + \frac{\pi^d \frac{\partial^2 \pi^d}{\partial x_j^2} - \left(\frac{\partial \pi^d}{\partial x_j}\right)^2}{(\pi^d)^2}\right\} \tag{3.108}$$

$$= \frac{1}{(d-1)}\sum_{j\neq i}\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}. \tag{3.109}$$

Next, we show $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2\right]$ converges to 0. To prove this, consider writing $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2\right]$ as sum of $(d-1)^2$ terms

$$\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2\right] = \frac{1}{(d-1)^2}\int \sup_i \sum_{j\neq i}\sum_{k\neq i}\left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2 \pi^d}{\partial x_k^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d \tag{3.110}$$

$$\leq \frac{1}{(d-1)^2}\sum_{j=1}^{d}\sum_{k=1}^{d}\int\left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2 \pi^d}{\partial x_k^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d - \frac{2}{(d-1)^2}\int \inf_i \sum_{j\neq i}\left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d \tag{3.111}$$

$$= \frac{1}{(d-1)^2}\sum_{j=1}^{d}\sum_{k=1}^{d}\int\left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2 \pi^d}{\partial x_k^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d + o(1), \tag{3.112}$$

where the last equality follows from

$$\frac{2}{(d-1)^2}\int \inf_i \sum_{j\neq i}\left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d \geq \frac{2}{(d-1)^2}\int \inf_i \sum_{j\neq i}\left(\frac{\partial^2 \log \pi_{-i}}{\partial x_j^2}\right)\pi^d \mathrm{d}x^d \tag{3.113}$$

$$= \frac{2}{(d-1)^2} o(d\sqrt{d/l_d}) = o(1). \tag{3.114}$$

When $|j-k| \geq l_d$, by Assumption A2, we have

$$\int \left(\frac{\partial^2 \pi^d}{\partial x_j^2}\frac{1}{\pi^d}\right)\left(\frac{\partial^2 \pi^d}{\partial x_k^2}\frac{1}{\pi^d}\right)\pi^d \mathrm{d}x^d \tag{3.115}$$

$$= \int \left(\frac{\partial^2 \pi^d}{\partial x_j^2}\right)\left(\frac{\partial^2 \pi^d}{\partial x_k^2}\right)\frac{1}{\pi^d}\mathrm{d}x^d \tag{3.116}$$

$$= \int \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_j^2}\right)\left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_k^2}\right)\frac{1}{\pi_{j,k|-j-k}}\pi_{-j-k}\mathrm{d}x_{-j-k}\mathrm{d}x_j\mathrm{d}x_k \tag{3.117}$$

$$\leq \int \left[\sup_{x^d \in F_d} \int \left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_j^2}\right)\left(\frac{\partial^2 \pi_{j,k|-j-k}}{\partial x_k^2}\right)\frac{1}{\pi_{j,k|-j-k}}\mathrm{d}x_j\mathrm{d}x_k\right]\pi_{-j-k}\mathrm{d}x_{-j-k} \tag{3.118}$$

$$\to 0. \tag{3.119}$$

This implies $\mathbb{E}\left[\sup_i (R_d^{(i)} + S_d^{(i)})^2\right] = \frac{\mathcal{O}(d\,l_d)+(d-l_d)^2 o(1)}{(d-1)^2} + o(1) \to 0$. Therefore, uniformly over i, $R_d^{(i)} + S_d^{(i)} \to 0$ in probability, then there exists a sequence $\{F_d'\}$ such that $\mathbb{P}(R_d^{(i)} + S_d^{(i)} \in F_d', \forall i) \to 1$ and the following holds

$$\sup_i \sup_{x^d \in F_d'} \left|R_d^{(i)} + S_d^{(i)}\right| \to 0. \tag{3.120}$$

Proof of Lemma 3.4.6

Note that Assumption A4 implies

$$\sup_{i\in\{1,\ldots,d\}} \sup_{x^d \in F_d^+} \frac{\partial}{\partial x_i}\log \pi^d(x^d) = o\left(d^{1/2}\right). \tag{3.121}$$

Then, by the definitions of $R_d^{(i)}$ and $I_d(x^d)$, we have

$$R_d^{(i)} - I_d(x^d) = \frac{1}{d-1}\sum_{j\neq i}\left(\frac{\partial \log \pi_{-i}(x_{-i}\,|\,x_i)}{\partial x_j}\right)^2 - \frac{1}{d}\sum_{j=1}^d\left(\frac{\partial}{\partial x_j}\log \pi^d(x^d)\right)^2 \tag{3.122}$$

$$= \frac{1}{d-1}\sum_{j\neq i}\left(\frac{\partial \log \pi^d(x^d)}{\partial x_j}\right)^2 - \frac{1}{d}\sum_{j=1}^d\left(\frac{\partial}{\partial x_j}\log \pi^d(x^d)\right)^2 \tag{3.123}$$

$$= \frac{1}{d}R_d^{(i)} - \frac{1}{d}\left(\frac{\partial}{\partial x_i}\log \pi^d(x^d)\right)^2 \to 0. \tag{3.124}$$

Proof of Lemma 3.4.7

Recall that we have shown

$$\hat{M}_{x^d}^{(i)}(Y_i) = \Phi\left(R_d^{(i)}(Y_i)^{-1/2}\left(\ell^{-1}\log\frac{\pi^d(y^d(i))}{\pi^d(x^d)} - \ell R_d^{(i)}(Y_i)/2\right)\right) \tag{3.125}$$

$$+ \exp\left(\log\frac{\pi^d(y^d(i))}{\pi^d(x^d)}\right)\Phi\left(-\ell R_d^{(i)}(Y_i)^{1/2}/2 - \log\frac{\pi^d(y^d(i))}{\pi^d(x^d)}R_d^{(i)}(Y_i)^{-1/2}\ell^{-1}\right). \tag{3.126}$$

For notational simplicity, we omit the index i and write $R_d^{(i)}$ by R_d. To simplify the derivation, we note that $\hat{M}_{x^d}^{(i)}(y)$ has the following form

$$M(y) = \Phi\left(f(y)g(y) - \frac{1}{2}f^{-1}(y)\right) + \exp(g(y))\Phi\left(-\frac{1}{2}f^{-1}(y) - f(y)g(y)\right), \tag{3.127}$$

where $f^{-1}(y) := \ell R_d^{1/2}(y)$ and $g(y) = \log \pi^d(y^d(i)) - \log \pi^d(x^d)$. Taking the derivative with respect to y, we get

$$\begin{aligned}
\frac{\mathrm{d}M(y)}{\mathrm{d}y} &= \Phi'(fg - f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2) && (3.128)\\
&\quad + \exp(g)\Phi'(-f^{-1}/2 - fg)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) && (3.129)\\
&\quad + \exp(g)\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi(-fg - f^{-1}/2) && (3.130)\\
&\leq \|\Phi'\|_\infty \left|\frac{\mathrm{d}f}{\mathrm{d}y}g + \frac{\mathrm{d}g}{\mathrm{d}y}f - \frac{1}{2}\frac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right| && (3.131)\\
&\quad + \exp(g)\|\Phi'\|_\infty \left|\frac{\mathrm{d}f}{\mathrm{d}y}g + \frac{\mathrm{d}g}{\mathrm{d}y}f + \frac{1}{2}\frac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right| && (3.132)\\
&\quad + \exp(g)\left|\frac{\mathrm{d}g}{\mathrm{d}y}\right|\|\Phi\|_\infty && (3.133)
\end{aligned}$$

Note that both Φ and Φ' are bounded functions. It then suffices to show

$$\begin{aligned}
&\exp(g)\left|\frac{\mathrm{d}g}{\mathrm{d}y}\right| = o(d^{1/2}), \quad \exp(g)\left|\frac{\mathrm{d}f}{\mathrm{d}y}g\right| = o(d^{1/2}), && (3.134)\\
&\exp(g)\left|\frac{\mathrm{d}g}{\mathrm{d}y}f\right| = o(d^{1/2}), \quad \exp(g)\left|\frac{\mathrm{d}f^{-1}}{\mathrm{d}y}\right| = o(d^{1/2}). && (3.135)
\end{aligned}$$

Observing that $\frac{\mathrm{d}f^{-1}}{\mathrm{d}y} = \frac{1}{2}\ell R_d'/R_d^{1/2}$ and $\frac{\mathrm{d}f}{\mathrm{d}y} = -\frac{1}{2\ell}\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}$, if we can show

$$\sup_{i\in\{1,\ldots,d\}} \frac{\mathrm{d}R_d^{(i)}(y)}{\mathrm{d}y}\frac{1}{[R_d^{(i)}(y)]^{1/2}} = o(1), \tag{3.136}$$

then we can get $\frac{\mathrm{d}f^{-1}}{\mathrm{d}y} = o(1)$ and $\frac{\mathrm{d}f}{\mathrm{d}y} = o(1/R_d)$. Using $R_d^{(i)} \to I_d(x^d)$ from Section 3.4.2, it suffices to show

$$\begin{aligned}
&\left(\sup_{x^d\in F_d^+}\pi^d(x^d)\right)\left(\sup_i \sup_{x^d\in F_d^{(i)}} \frac{\partial \log \pi^d(x^d)}{\partial x_i}\right) = o(d^{1/2}), && (3.137)\\
&\left(\sup_{x^d\in F_d^+}\pi^d(x^d)\right)\left(\sup_{x^d\in F_d^+} \left|\log(\pi^d(x^d))/I_d(x^d)\right|\right) = o(d^{1/2}), && (3.138)\\
&\left(\sup_{x^d\in F_d^+}\pi^d(x^d)\right)\left(\sup_i \sup_{x^d\in F_d^{(i)}} \left|\frac{\partial \log \pi^d(x^d)}{\partial x_i}/\sqrt{I_d(x^d)}\right|\right) = o(d^{1/2}). && (3.139)
\end{aligned}$$

One can easily verify that the above equations hold under Assumption A4.

Finally, we complete the proof by showing Eq. (3.136). Recall that

$$R_d^{(i)}(y) = \frac{1}{d-1}\sum_{j\neq i}\left(\frac{\partial \log \pi_{-i}(x_{-i}\,|\,x_i = y)}{\partial x_j}\right)^2. \tag{3.140}$$

For notational simplicity, we write

$$R_d^{(i)}(y) = \frac{1}{d-1}\sum_{j\neq i} f_j^2(y), \tag{3.141}$$

where $f_j(y) := \frac{\partial \log \pi_{-i}(x_{-i}\,|\,x_i=y)}{\partial x_j}$. Then, by Cauchy–Schwartz inequality

$$\frac{\partial R_d^{(i)}(y)}{\mathrm{d}y} = \frac{2}{d-1}\sum_{j\neq i} f_j(y) f_j'(y) \le \frac{2}{d-1}\sqrt{\sum_{j\neq i} f_j^2(y)\sum_{j\neq i}|f_j'(y)|^2}. \tag{3.142}$$

Note that by A1, if $|i-j| > l_d$ then $f_j'(y) \le \sup_{x^d\in F_d}\frac{\partial^2 \log \pi^d(x^d)}{\partial x_i \partial x_j} \to 0$. Hence, we have

$$\sup_{i\in\{1,\ldots,d\}}\frac{\mathrm{d}R_d^{(i)}(y)}{\mathrm{d}y}\frac{1}{[R_d^{(i)}(y)]^{1/2}} \le \sup_i \frac{\frac{2}{d-1}\sqrt{\sum_{j\neq i} f_j^2(y)\sum_{j\neq i}|f_j'(y)|^2}}{\sqrt{\frac{1}{d-1}\sum_{j\neq i} f_j^2(y)}} \tag{3.143}$$

$$= 2\sup_i\sqrt{\frac{1}{d-1}\sum_{j\neq i}|f_j'(y)|^2} \le 2\sqrt{\frac{1}{d-1}\sum_{j=1}^{d}|f_j'(y)|^2} = o\left(\sqrt{\frac{l_d}{d}(\sqrt{d/l_d})^2}\right) = o(1). \tag{3.144}$$

3.4.3 Proof of Theorem 3.2.19

Similar to Section 3.4.1, we assume the coordinates are linear ordered for simplicity. The proof follows the framework of [RGG97] using the generator approach [EK86].

Define the (discrete time) generator of x^d by

$$(G_d f)(x^d) := d\mathbb{E}_{Y^d}\left\{[f(Y^d) - f(x^d)]\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)}\right)\right\}, \tag{3.145}$$

for any function f for which this definition makes sense. In the Skorokhod topology, it doesn't cause any problem to treat G_d as a continuous time generator. We shall restrict attention to test functions such that $f(x^d) = f(x_1)$. We show uniform convergence of G_d to G, the generator of the limiting (one-dimensional) Langevin diffusion, for a suitable large class of real-valued functions f, where, for some fixed function $h(\ell)$,

$$(Gf)(x_1) := h(\ell)\left\{\frac{1}{2}f''(x_1) + \frac{1}{2}[(\log\tilde{\pi})'(x_1)]\,f'(x_1)\right\}, \tag{3.146}$$

in which $\tilde{\pi}$ is a one-dimensional density of the first coordinate of π^d. Since we have assumed in A6 that $(\log\tilde{\pi})'$ is Lipschitz, by [EK86, Thm 2.1 in Ch.8], a core for the generator has domain C_c^∞, which is the class of continuous functions with compact support such that all orders of derivatives exist. This enable us to restrict attentions to functions $f_c \in C_c^\infty$ such that $f_c(x^d) = f_c(x_1)$.

Note that using Assumption A2+, and the assumption $\pi^d(F_d^c) = \mathcal{O}(d^{-1-\delta})$, following the arguments in

the proof of Lemma 3.4.3 we can get a stronger version of Lemma 3.4.3 for $F_d' := \{x^d : \sup_i |R_d^{(i)} + S_d^{(i)}| \leq d^{-\delta}\}$. Then using a union bound yields

$$\mathbb{P}(X^d(\lfloor ds \rfloor) \notin F_d \cap F_d', \exists 0 \leq s \leq t) \to 0. \tag{3.147}$$

Therefore, for any fixed t, if $d \to \infty$ then the probability of all $X^d(\lfloor ds \rfloor), 0 \leq s \leq t$ are in $F_d \cap F_d'$ goes to 1. Since $F_d \cap F_d' \subseteq F_d^+ \cap F_d' \subseteq F_d^+$, it suffices to consider $x^d \in F_d^+$.

Note that $Y^d \sim \mathcal{N}(x^d, \frac{\ell^2}{d-1} I)$, we can write

$$(G_d f_c)(x^d) = d\mathbb{E}_{Y_1}\left\{[f_c(Y_1) - f_c(x_1)]\mathbb{E}_{Y_{-1}}\left[1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)}\right]\right\}, \tag{3.148}$$

where $\mathbb{E}_{Y_{-1}}[\cdot]$ is short for $\mathbb{E}_{Y_2,\ldots,Y_d \,|\, Y_1}[\cdot]$ and π^d denotes the target distribution in d-dimension. The goal is then to prove $(G_d f_c)$ converges to (Gf_c).

Recall the definition Eq. (3.56), we omit the index to write $M_{x^d}^{(1)}$ as M_{x^d}, which is defined by

$$M_{x^d}(Y_1) = \mathbb{E}_{Y_{-1}}\left(1 \wedge \frac{\pi^d(Y^d)}{\pi^d(x^d)}\right). \tag{3.149}$$

Then we have previously shown in Eq. (3.79) that $M_{x^d}(Y_1)$ can be approximated by

$$\hat{M}_{x^d}(Y_1) = \Phi\left(R_d(Y_1)^{-1/2}\left(\ell^{-1}\log\frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)} - \ell R_d(Y_1)/2\right)\right) \tag{3.150}$$

$$+ \exp\left(\log\frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)}\right)\Phi\left(-\ell R_d(Y_1)^{1/2}/2 - \log\frac{\pi^d(Y_1, x_{-1})}{\pi^d(x^d)} R_d(Y_1)^{-1/2}\ell^{-1}\right) \tag{3.151}$$

For $x^d \in F_d^+$, some properties of $\hat{M}_{x^d}$ is given as follows.

Lemma 3.4.9. *For $\hat{M}_{x^d}$, we have*

$$\hat{M}_{x^d}(x_1) = 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right), \tag{3.152}$$

$$\hat{M}'_{x^d}(x_1) = \Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{\mathrm{d}[\log\pi_1(x) + \log\pi_{-1}(x_{-1}\,|x)]}{\mathrm{d}x}(x_1) + o(1), \tag{3.153}$$

$$\hat{M}'_{x^d}(x_1) = o(d^{1/2}), \quad \sup_{x^d \in F_d^+} \hat{M}''_{x^d} = o(d^{1/2}). \tag{3.154}$$

Proof. See Section 3.4.4. □

Since $f_c(Y_1) - f_c(x_1)$ is bounded, it suffices to show

$$\mathbb{E}_{Y_1}\left\{d[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} \to (Gf_c)(x_1). \tag{3.155}$$

Now using mean value theorem and Taylor expansion of $\mathbb{E}_{Y_1}\left\{[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\}$ at $(Y_1 - x_1)$

yields

$$[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1) \tag{3.156}$$
$$= \left[f_c'(x_1)(Y_1 - x_1) + \frac{1}{2}f_c''(x_1)(Y_1 - x_1)^2 + K(Y_1 - x_1)^3\right] \tag{3.157}$$
$$\cdot \left[\hat{M}_{x^d}(x_1) + \hat{M}_{x^d}'(x_1)(Y_1 - x_1) + \frac{1}{2}\hat{M}_{x^d}''(x')(Y_1 - x_1)^2\right] \tag{3.158}$$
$$= f_c'(x_1)\hat{M}_{x^d}(x_1)(Y_1 - x_1) + \left[\frac{1}{2}f_c''(x_1)\hat{M}_{x^d}(x_1) + f_c'(x_1)\hat{M}_{x^d}'(x_1)\right](Y_1 - x_1)^2 \tag{3.159}$$
$$+ \left[K\hat{M}_{x^d}(x_1) + \frac{1}{2}f_c''(x_1)\hat{M}_{x^d}'(x_1) + \frac{1}{2}\hat{M}_{x^d}''(x')f_c'(x_1)\right](Y_1 - x_1)^3 \tag{3.160}$$
$$+ \left[\frac{1}{4}\hat{M}_{x^d}''(x')f_c''(x_1) + K\hat{M}_{x^d}'(x_1)\right](Y_1 - x_1)^4 + \frac{1}{2}\hat{M}_{x^d}''(x')K(Y_1 - x_1)^5, \tag{3.161}$$

where K is a constant since f_c has bounded third derivative. Note that both $f_c'(x_1)$ and $f_c''(x_1)$ are bounded as well. Therefore, taking expectation over Y_1 and using $\hat{M}_{x^d}'(x_1) = o(d^{1/2}), \sup_{x^d}\hat{M}_{x^d}'' = o(d^{1/2})$ in Lemma 3.4.9, we have

$$\mathbb{E}_{Y_1}\left\{[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} = \left[\frac{1}{2}f_c''(x_1)\hat{M}_{x^d}(x_1) + f_c'(x_1)\hat{M}_{x^d}'(x_1)\right]\frac{\ell^2}{d-1} + o(d^{-1}). \tag{3.162}$$

Finally, by Assumption A6, we have

$$f_c'(x_1)\hat{M}_{x^d}'(x_1) + \frac{1}{2}f_c''(x_1)\hat{M}_{x^d}(x_1) \tag{3.163}$$
$$= 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}[\log\pi_1(x) + \log\pi_{-1}(x_{-1}\,|x)]}{\mathrm{d}x}(x_1)\right) \tag{3.164}$$
$$= 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\pi_{1|-1}(x\,|\,x_{-1})}{\mathrm{d}x}(x_1)\right) \tag{3.165}$$
$$\to 2\Phi\left(-\frac{\ell I(x^d)^{1/2}}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\tilde{\pi}(x)}{\mathrm{d}x}(x_1)\right) \tag{3.166}$$
$$\to 2\Phi\left(-\frac{\ell\bar{I}^{1/2}}{2}\right)\left(\frac{1}{2}f_c''(x_1) + \frac{1}{2}f_c'(x_1)\frac{\mathrm{d}\log\tilde{\pi}(x)}{\mathrm{d}x}(x_1)\right), \tag{3.167}$$

which implies that $\mathbb{E}_{Y_1}\left\{d[f_c(Y_1) - f_c(x_1)]\hat{M}_{x^d}(Y_1)\right\} \to (Gf_c)(x_1)$ where $h(\ell) := 2\ell^2\Phi(-\ell\sqrt{\bar{I}}/2)$.

3.4.4 Proof of Lemma 3.4.9

The proof is quite tedious. In order to simplify the notations, we first introduce the following lemma.

Lemma 3.4.10. *For the function $M(y)$ defined by*

$$M(y) = \Phi\left(f(y)g(y) - \frac{1}{2}f^{-1}(y)\right) + e^{g(y)}\Phi\left(-\frac{1}{2}f^{-1}(y) - f(y)g(y)\right), \tag{3.168}$$

we have

$$\frac{\mathrm{d}M(y)}{\mathrm{d}y} = \Phi'(fg - f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2) \tag{3.169}$$

$$+ e^g\Phi'(-f^{-1}/2 - fg)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \tag{3.170}$$

$$+ e^g\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi(-fg - f^{-1}/2). \tag{3.171}$$

$$\frac{\mathrm{d}^2M(y)}{\mathrm{d}y^2} = \Phi''(fg - f^{-1}/2)\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2 + \Phi'(fg - f^{-1}/2)\frac{\mathrm{d}^2}{\mathrm{d}y^2}(fg - f^{-1}/2) \tag{3.172}$$

$$+ e^g\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi'(-f^{-1}/2 - fg)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \tag{3.173}$$

$$+ e^g\left\{\Phi''(-fg - f^{-1}/2)\left[\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2)\right]^2 + \Phi'(-fg - f^{-1}/2)\frac{\mathrm{d}^2}{\mathrm{d}y^2}(-fg - f^{-1}/2)\right\} \tag{3.174}$$

$$+ e^g\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi'(-fg - f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \tag{3.175}$$

$$+ \Phi(-fg - f^{-1}/2)\left[e^g\left(\frac{\mathrm{d}^2}{\mathrm{d}y^2}g\right) + e^g\left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)^2\right]. \tag{3.176}$$

Furthermore, if $g(x_1) = 0$*, then we have*

$$\frac{\mathrm{d}M(y)}{\mathrm{d}y}(x_1) = \left(\Phi'(-f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right. \tag{3.177}$$

$$+\Phi'(-f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2) \tag{3.178}$$

$$\left. + \left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi(-f^{-1}/2)\right)(x_1) \tag{3.179}$$

$$= \left(\Phi'(-f^{-1}/2)\frac{\mathrm{d}}{\mathrm{d}y}(-f^{-1}) + \left(\frac{\mathrm{d}}{\mathrm{d}y}g\right)\Phi(-f^{-1}/2)\right)(x_1) \tag{3.180}$$

$$= -\Phi'\left(-\frac{f^{-1}(x_1)}{2}\right)\frac{\mathrm{d}f^{-1}(y)}{\mathrm{d}y}(x_1) + \frac{\mathrm{d}g(y)}{\mathrm{d}y}(x_1)\Phi\left(-\frac{f^{-1}(x_1)}{2}\right). \tag{3.181}$$

Remark 3.4.11. Let $g(y) = \log\frac{\pi^d(y,x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$ then $\hat{M}_{x^d}(y) = M(y)$.

Now substituting $g(y) = \log\frac{\pi^d(y,x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$ to Lemma 3.4.10, we have

$$\hat{M}_{x^d}(x_1) = 2\Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right), \tag{3.182}$$

and

$$\hat{M}'_{x^d}(x_1) = \frac{\mathrm{d}\hat{M}_{x^d}(y)}{\mathrm{d}y}(x_1) \tag{3.183}$$

$$= \Phi\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{\mathrm{d}[\log\pi_1(x) + \log\pi_{-1}(x_{-1}\,|x)]}{\mathrm{d}x}(x_1) \tag{3.184}$$

$$- \Phi'\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{\ell}{2R_d^{1/2}(x_1)}R'_d(x_1). \tag{3.185}$$

Since Φ' is bounded and by Eq. (3.136), $R'_d(x_1)/R_d^{1/2}(x_1) \to 0$, therefore

$$\Phi'\left(-\frac{\ell R_d^{1/2}(x_1)}{2}\right)\frac{\ell}{2R_d^{1/2}(x_1)}R'_d(x_1) = o(1). \tag{3.186}$$

Also, $\hat{M}'_{x^d}(x_1) = o(d^{1/2})$ since $\frac{\partial\log\pi^d}{\partial x_i} = \mathcal{O}(d^{\alpha/2}) = o(d^{1/2})$.

Now we prove $\sup_{x^d}\hat{M}''_{x^d} = o(d^{1/2})$. For simplicity, we keep the notations of f and g (recall that $g(y) = \log\frac{\pi^d(y,x_{-1})}{\pi^d(x^d)}$ and $f^{-1}(y) = \ell R_d^{1/2}(y)$) and use the results in Section 3.4.2. Since Φ, Φ', Φ'' are bounded, it suffices to bound all the following terms to be $o(d^{1/2})$:

$$\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2, \quad \frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2), \quad \exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2), \tag{3.187}$$

$$\exp(g)\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2, \quad \exp(g)\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2), \quad \exp(g)\left(\frac{\mathrm{d}^2 g}{\mathrm{d}y^2}\right), \quad \exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)^2. \tag{3.188}$$

Next, we show that most of them can be verified using Assumption A4+, and the results in Section 3.4.2:

$$\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2 = \mathcal{O}\left[\left(\sup_{x^d\in F_d^+}\log\pi^d(x^d)\mathcal{O}(d^{\alpha/4}) + \sup_{x^d\in F_d^+}\frac{\partial\log\pi^d}{\partial x_1}\right)^2\right] \tag{3.189}$$

$$= \mathcal{O}\left[(d^{\alpha/4}\log d + d^{\alpha/2})^2\right] = o(d^{1/2}), \tag{3.190}$$

$$\left|e^g(\frac{\mathrm{d}g}{\mathrm{d}y})\frac{\mathrm{d}}{\mathrm{d}y}(-fg - f^{-1}/2)\right| = \mathcal{O}\left[\sup_{x^d\in F_d^+}\pi^d(x^d)\sup_{x^d\in F_d^+}\frac{\partial\log\pi^d}{\partial x_1}\left(d^{\alpha/4}\log d + d^{\alpha/2}\right)\right] \tag{3.191}$$

$$= o(d^{1/2-\alpha}d^{\alpha/2}(d^{\alpha/4}\log d + d^{\alpha/2})) = o(d^{1/2}), \tag{3.192}$$

$$\left|\exp(g)\left[\frac{\mathrm{d}}{\mathrm{d}y}(fg - f^{-1}/2)\right]^2\right| = o(d^{1/2-\alpha}d^{\alpha}) = o(d^{1/2}), \tag{3.193}$$

$$\left|\exp(g)\left(\frac{\mathrm{d}^2 g}{\mathrm{d}y^2}\right)\right| = \mathcal{O}\left[\sup_{x^d\in F_d^+}\pi^d(x^d)\sup_{x^d\in F_d^+}\frac{\partial^2\log\pi^d}{\partial x_1^2}\right] = o(d^{1/2-\alpha})\mathcal{O}(d^{\alpha}) = o(d^{1/2}), \tag{3.194}$$

$$\left|\exp(g)\left(\frac{\mathrm{d}g}{\mathrm{d}y}\right)^2\right| = \mathcal{O}\left[\sup_{x^d\in F_d^+}\pi^d(x^d)\sup_{x^d\in F_d^+}\left(\frac{\partial\log\pi^d}{\partial x_1^2}\right)^2\right] = o(d^{1/2-\alpha})\mathcal{O}(d^{\alpha/2})^2 = o(d^{1/2}). \tag{3.195}$$

The only terms left are $\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2)$ and $\exp(g)\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2)$. Therefore, it suffices to show

$$\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2) = \mathcal{O}(d^{\alpha}). \tag{3.196}$$

Note that

$$\frac{\mathrm{d}^2}{\mathrm{d}y}(fg - f^{-1}/2) = \frac{\mathrm{d}}{\mathrm{d}y}(f'g + g'f - \frac{1}{2}\mathrm{d}f^{-1}) \tag{3.197}$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}\left[\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g + \frac{1}{R_d^{1/2}}g' - \frac{1}{2}\frac{R_d'}{R_d^{1/2}}\right] \tag{3.198}$$

$$= \frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g' + \left(\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}\right)'g + \frac{1}{R_d^{1/2}}g'' + \left(\frac{1}{R_d^{1/2}}\right)'g' - \frac{1}{2}\left(\frac{R_d'}{R_d^{1/2}}\right)'. \tag{3.199}$$

Note that we have shown $R_d' = o(R_d^{1/2})$ in Section 3.4.2. Similarly, we also can show using Assumption A3+ that

$$R_d'' = \frac{1}{d-1}(\sum_{j\neq 1} f_j f_j')' = \frac{1}{d-1}\sum_{j\neq 1}(f_j')^2 + \frac{1}{d-1}\sum_{j\neq 1} f_j f_j'' \tag{3.200}$$

$$\leq \frac{1}{d-1}\sum_{j\neq 1}(f_j')^2 + \sqrt{\frac{1}{d-1}\sum_{j\neq 1} f_j^2}\sqrt{\frac{1}{d-1}\sum_{j\neq 1}(f_j'')^2} \tag{3.201}$$

$$= \mathcal{O}(l_d/d)o((\sqrt{d/l_d})^2) + o(R_d^{1/2}\sqrt{l_d/d(\sqrt{d/l_d})^2}) = o(R_d^{1/2}), \tag{3.202}$$

where $f_j(x) := \frac{\partial \log \pi_{-1}(x_{-1}\,|\,x_1=x)}{\partial x_j}$. Therefore $R_d'' = o(R_d^{1/2})$ as well. Finally, we can complete the proof by verifying Eq. (3.196) using Assumption A4+ as follows.

$$\left|\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}g'\right| = \mathcal{O}\left(\frac{1}{R_d}\right)o(1)\mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial \log \pi^d}{\partial x_1}\right) = \mathcal{O}(d^{\alpha/4})o(d^{\alpha/2}) = o(d^\alpha), \tag{3.203}$$

$$\left|\left(\frac{1}{R_d}\frac{R_d'}{R_d^{1/2}}\right)'g\right| = \mathcal{O}\left[\frac{R_d''R_d^{3/2} + 3/2(R_d')^2R_d^{1/2}}{R_d^3}g\right] = \mathcal{O}\left[\frac{1}{R_d^{3/2}}(R_d''g)\right] \tag{3.204}$$

$$= \mathcal{O}(d^{\alpha/4})o(1)\mathcal{O}(d^{\alpha/2}) = o(d^\alpha), \tag{3.205}$$

$$\left|\frac{1}{R_d^{1/2}}g''\right| = \mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial^2 \log \pi^d}{\partial x_1^2}\right) = o(d^\alpha), \tag{3.206}$$

$$\left|\left(\frac{1}{R_d^{1/2}}\right)'g'\right| = \mathcal{O}\left(\frac{1}{2}\frac{1}{R_d^{3/2}}R_d'g'\right) = o(1/R_d)\mathcal{O}\left(\sup_{x^d\in F_d^+}\frac{\partial \log \pi^d}{\partial x_1}\right) = \mathcal{O}(d^{\alpha/4})o(d^{\alpha/2}) = o(d^\alpha), \tag{3.207}$$

$$\left|\left(\frac{R_d'}{R_d^{1/2}}\right)'\right| = \left|\frac{R_d''R_d^{1/2} - \frac{1}{2}(R_d')^2\frac{1}{R_d^{1/2}}}{R_d}\right| = \mathcal{O}\left(R_d''/R_d^{1/2}\right) = o(1) = o(d^\alpha). \tag{3.208}$$

3.4.5 Proof of Theorem 3.2.21

We follow the same approach as in the proof of [RR16, Proposition 3]. The idea is to follow the proof of Theorem 3.2.19 except in the proof of Eq. (3.79), we need a stronger version of Lemma 3.4.3 to determine the sequence of "typical sets" $\{F_d'\}$.

Given fixed time t, considering the sequence of "typical sets" $\{F_d'\}$ defined by

$$F_d' := \{x^d : |R_d + S_d| \leq d^{-\delta}\}, \tag{3.209}$$

where $\delta > 0$ and we used R_d and S_d to denote $R_d^{(1)}$ and $S_d^{(1)}$ for simplicity. We need to guarantee that

when d is large enough, we always have $X^d(\lfloor ds \rfloor) \in F_d \cap F'_d, \forall 0 \leq s \leq t$ and this happens for almost all starting state $X_1^d(0) = x$. That is, defining

$$p(d,x) := \mathbb{P}(X(\lfloor ds \rfloor) \notin F_d \cap F'_d, \exists 0 \leq s \leq t \,|\, X_1^d(0) = x), \tag{3.210}$$

letting π_1 denote the marginal stationary distribution for the first coordinate, we want to show that for any given $\epsilon > 0$, as $d \to \infty$

$$\mathbb{P}_{x\sim\pi_1}[p(d,x) \geq \epsilon, \text{infinite often}] = 0. \tag{3.211}$$

We prove it using Borel–Cantelli Lemma. Note that the application of Borel–Cantelli lemma is valid since we have assumed all of the processes are jointly defined on the same probability space as independent processes. First, note that

$$\mathbb{E}_{x\sim\pi_1}[p(d,x)] = dt\mathbb{P}_{\pi^d}((F_d \cap F'_d)^c) = dt\mathbb{P}_{\pi^d}(F_d^c \cup (F'_d)^c) \leq dt\mathbb{P}_{\pi^d}(F_d^c) + dt\mathbb{P}_{\pi^d}((F'_d)^c). \tag{3.212}$$

For any given $\epsilon > 0$, we have

$$\sum_{d=2}^{\infty} \mathbb{P}(p(x,d) \geq \epsilon) \leq \sum_{d=2}^{\infty} \frac{\mathbb{E}_{x\sim\pi_1}[p(d,x)]}{\epsilon} \tag{3.213}$$

$$\leq \frac{dt}{\epsilon} \sum_{d=2}^{\infty} \mathbb{P}_{\pi^d}(|R_d + S_d| > d^{-\delta}) + \frac{dt}{\epsilon} \sum_{d=2}^{\infty} \mathbb{P}(X^d \notin F_d). \tag{3.214}$$

By $\pi^d(F_d^c) = \mathcal{O}(d^{-2-\delta})$, we have $dt \sum_{d=2}^{\infty} \mathbb{P}(X^d \notin F_d) < \infty$. Now in order to use Borel–Cantelli Lemma, the condition we need is that for some number of moments m such that

$$\mathbb{P}_{\pi^d}(|R_d + S_d| > d^{-\delta}) \leq \frac{\mathbb{E}|R_d + S_d|^m}{d^{-m\delta}} = d^{m\delta}\mathbb{E}|R_d + S_d|^m = \mathcal{O}(d^{-2-\delta}), \tag{3.215}$$

which leads to $\sum_{d=2}^{\infty} \mathbb{P}(p(x,d) \geq \epsilon) < \infty$. In order to obtain non-trivial conditions, we let $m = 5$ and Assumption A2++ implies $\mathbb{E}|R_d + S_d|^5 = \mathcal{O}(d^{-2-6\delta})$. We can then use this sequence of typical sets $\{F'_d\}$ in the proof of Theorem 3.2.19 to replace the sequence of $\{F'_d\}$ used in Lemma 3.4.3. The residual proof follows the same as Theorem 3.2.19.

3.4.6 Proof of Proposition 3.3.6

Note that we have the number of parameters $d = n^2 + n + 2$ in this example. The target distribution (i.e. the posterior distribution) satisfies

$$\begin{aligned}\pi^d(x^d) &= \mathbb{P}(x^d \mid \{Y_{ij}\}_{i,j=1}^n) \\ &\propto \frac{b^a}{\Gamma(a)} A^{-a-1} e^{-b/A} \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi A}} e^{-\frac{(\mu_j - \nu)^2}{2A}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi V}} e^{-\frac{(\theta_{ij} - \mu_j)^2}{2V}} \frac{1}{\sqrt{2\pi W}} e^{-\frac{(Y_{ij} - \theta_{ij})^2}{2W}}.\end{aligned} \tag{3.216}$$

Clearly, this model can be represented by the graphical model in Eq. (3.41). It can be easily checked that the maximum number cliques any coordinate belongs to is $n + 1$ and the cardinality of cliques is bounded by constant 2, so $\sup_k |C_k| = o(d/m_d) = o(n)$. Furthermore, the target distribution clearly satisfies "flat tail" condition required by Proposition 3.3.3 since all the conditional distributions are

standard distributions. Therefore, the first equation in A1, the first equation in A3, and A2 hold by Proposition 3.3.3.

Next, we verify A5 using Proposition 3.3.4. Note that this model can be represented by the graphical model in Eq. (3.42) using $K = 3$ layers. In order to check the conditions in Proposition 3.3.4, note that

$$\log \pi^d \propto (-a - 1 - \frac{n}{2}) \log A - \frac{b}{A} - \frac{\sum_j (\mu_j - \nu)^2}{2A} - \frac{\sum_{i,j} (\theta_{ij} - \mu_j)^2}{2V} - \frac{\sum_{i,j} (Y_{ij} - \theta_{ij})^2}{2W}. \tag{3.217}$$

Observing that, under $X^d = (\nu, A, \{\mu_j\}_{j=1}^n, \{\theta_{ij}\}_{i,j=1}^n) \sim \pi^d$, we have

$$\theta_{ij} \mid Y_{ij}, \mu_j \quad \sim^{\text{indep.}} \mathcal{N}\left(\frac{W\mu_j + VY_{ij}}{W+V}, \frac{VW}{W+V}\right), \quad i, j \in \{1, \ldots, n\}, \tag{3.218}$$

$$\mu_j \mid \sum_i \theta_{ij}, \nu, A \sim^{\text{indep.}} \mathcal{N}\left(\frac{\sum_i A\theta_{ij} + V\nu}{nA+V}, \frac{AV}{nA+V}\right), \quad i \in \{1, \ldots, n\}, \tag{3.219}$$

$$\nu \mid \bar{\mu}, A \sim \mathcal{N}\left(\bar{\mu}, \frac{A}{n}\right), \tag{3.220}$$

$$A \mid \{\mu_j\}, \nu \sim \textbf{IG}\left(a + \frac{n}{2}, b + \frac{1}{2}\sum_j (\mu_j - \nu)^2\right). \tag{3.221}$$

Therefore, we have

$$\left|\frac{\partial \log \pi^d}{\partial A}\right| = \left|\frac{b + \frac{1}{2}\sum_j (\mu_j - \nu)^2}{A^2} - \frac{a + 1 + \frac{n}{2}}{A}\right| = \mathcal{O}_{\mathbb{P}}(d^{1/2}). \tag{3.222}$$

since $\frac{a+1+n/2}{A} \to_{\mathbb{P}} \frac{a+1+n/2}{A_0} = \mathcal{O}(d^{1/2})$ and $\sum_j (\mu_j - \nu)^2 \to_{\mathbb{P}} \sum_j (\mu_j - \bar{\mu})^2 + \frac{A_0}{n} = \mathcal{O}_{\mathbb{P}}(d^{1/2})$. Other coordinates can also be verified, which are shown as follows.

$$\left(\frac{\partial \log \pi^d}{\partial \nu}\right)^2 = \left(\frac{n(\bar{\mu} - \nu)}{A}\right)^2 = \mathcal{O}_{\mathbb{P}}\left(\frac{n}{A}\right) = \mathcal{O}_{\mathbb{P}}(d/n), \tag{3.223}$$

$$\left(\frac{\partial \log \pi^d}{\partial \mu_j}\right)^2 = \left(\frac{\sum_i (\theta_{ij} - \mu_j)}{V} - \frac{\mu_j - \nu}{A}\right)^2 = (nA + V)^2 \left(\frac{A\sum_i \theta_{ij} + V\nu}{nA+V} - \mu_j\right)^2 \tag{3.224}$$

$$= \mathcal{O}_{\mathbb{P}}\left[(nA+V)^2 \frac{AV}{nA+V}\right] = \mathcal{O}_{\mathbb{P}}(d/n), \tag{3.225}$$

$$\left(\frac{\partial \log \pi^d}{\partial \theta_{ij}}\right)^2 = \left(\frac{Y_{ij} - \theta_{ij}}{V} - \frac{\theta_{ij} - \mu_j}{W}\right)^2 = (W+V)^2 \left(\frac{VY_{ij} + W\mu_j}{W+V} - \theta_{ij}\right)^2 = \mathcal{O}_{\mathbb{P}}(d/n^2). \tag{3.226}$$

$$\tag{3.227}$$

Therefore, A5 holds by Proposition 3.3.4. Finally, all the other conditions in A1, A3, and A4 can be verified in a similar way as in Section 3.3.1 for Example 3.3.1.

Chapter 4

A Bayesian decision-theoretic analysis of Bayesian model misspecification

One of the hallmarks of Bayesian analysis is the use of "prior distributions". If we adopt the classical notion of a statistical model, where the data are assumed to be distributed according to one in a parametric family of probability distributions, the prior distribution is assumed to capture the statistician's prior knowledge and/or subjective beliefs about which parameters are most likely. The combination of the prior and model yields a joint distribution on the data and parameters. If a model is misspecified, it is often the case that no prior distribution yields a joint distribution that accurately represents the statistician's uncertainty. In the misspecified setting, the usual notion of a "subjective" prior may make no sense. Indeed, the statistician may believe that every available parameter setting should be assigned zero prior probability. As essentially every statistical model is misspecified, this raises the question: what is a prior? We focus on this setting of a misspecified model and the question of what prior distribution should be. Our solution is to view inference in pragmatic terms. Relative to one's beliefs, there is a *surrogate* prior that is most likely to produce the best answers.

In this work, we formalize the problem of *choosing* a (surrogate) prior as a Bayesian decision theory task, and develop theory for choosing optimal surrogate priors. The resulting framework, which we call *meta-Bayesian analysis*, gives (optimal surrogate) priors a pragmatic interpretation: relative to one's actual subjective beliefs, they lead to the best inference possible using the misspecified model. We discuss some early results on meta-Bayesian analysis in this work, which have some surprising consequences. For example, in violation of tradition Bayesian tenets, the optimal surrogate prior may depend on the loss function, on the number of data points you plan to observe, and on the number of predictions you expect to make. Furthermore, for general cases when the belief is a mixture of i.i.d. distributions, we show that under certain conditions, the belief on the "asymptotic locations" of posterior distributions is an asymptotically optimal prior. Finally, we study the i.i.d Bernoulli model relative to a general stationary belief on a binary sequence. We show that, under some conditions, the marginal prior belief on the limiting frequency of ones is asymptotically optimal when the number of predictions goes to infinity.

4.1 Related work on misspecification

There is a long history of work on inference and prediction under model misspecification. The study of Bayesian analysis under misspecification has its roots in the study of the asymptotic behavior of the maximum likelihood estimator (MLE). Building on work characterizing the consistency and asymptotic normality of the MLE in well-specified models [Wal49; LeC53], Berk [Ber66] showed that, asymptotically, a sequence of posterior distributions need not converge but will eventually be confined to the set of "pseudo-true" parameter values, i.e., the parameter values minimizing the Kullback–Liebler divergence with the distribution of the data. Later, Berk [Ber70] studied conditions under which a sequence of posterior distributions converges weakly to a degenerate distribution. Independently of [Ber66], Huber [Hub67] established general conditions implying the consistency and asymptotic normality of the MLE under model misspecification. (Huber's results actually apply beyond the MLE.) Simpler, though more restrictive, conditions were later given by White [Whi82]. The limit obtained by Huber agrees with that by Berk: The MLE ultimately stays within any neighborhood of the set of pseudo-true values. When the limit is unique, Huber also established the MLE's asymptotic covariance, now known as the "sandwich" covariance matrix.

There is a vast literature building on these foundational results. In the Bayesian context, the asymptotic normality of the posterior distribution under misspecification was studied by Chen [Che85] and Bunke and Milhaud [BM98]. Their work establishes that the posterior distribution is asymptotically normal, centered at the MLE, but with covariance equal to the reciprocal of the second derivative of the log likelihood function, not the sandwich covariance matrix. Royall and Tsou [RT03] show that the posterior distribution based on the adjusted (profile) likelihood function [Sta96] can be robust asymptotically, in the sense that it agrees with the posterior under the true model in the limit. Fushiki [Fus05] compares the Bayesian predictive distribution with the Bootstrap prediction obtained by applying Breiman's "bagging" method under the KL loss. It is shown that the Bootstrap prediction dominates the Bayesian prediction when model is misspecified. Müller [Mül13] shows that Bayesian inference about the pseudo-true parameter under squared error has lower frequentist risk asymptotically when the posterior is substituted by an artificial normal posterior centered at the MLE with sandwich covariance matrix. Other studies of artificial likelihoods and posteriors using the sandwich covariance matrix include [HW13] and [SD12]. These studies work with the asymptotic normal form of the posterior, where there is typically no longer a role played by the prior.

Lindsey [Lin99] discusses model misspecification within the Bayesian context. There, he questions the existence of a "true" model, and points out the contradiction between assigning probability one to a model at the outset and, at the same time, planning to check the model after seeing the data. He suggests that we should rather think in terms of models being appropriate simplifications, which are useful for detecting and understanding generalizable patterns in data, and for prediction. The type of patterns and the predictions, as well as the models to be used will depend on the questions to be answered. The validity of Bayesian probabilistic statements requires that the prior be personal and be specified before obtaining new information. However, Lindsey [Lin99] proposes neither an alternative meaning of priors, nor a method for choosing them.

Similar ideas appear in the discussion on specification and interpretation of nonparametric priors by Walker et al. [Wal+99]. An interesting example is given that, using the Bayesian nonparametric approach, one can ensure that the first two moments of the unknown function match those derived from a parametric model, which effectively creates a region where the function is thought to be located that is

the same for both parametric and nonparametric cases. The difference is that in the parametric case the shape of the unknown function is restricted whereas in the nonparametric case it is not. Gutierrez-Pena and Walker [GPW01] propose a method for model selection under model misspecification by working within a larger nonparametric model and taking a Bayesian predictive approach. Their approach involves performing posterior inference in a larger nonparametric model and then selecting the model whose posterior predictive distribution is closest according to some measure of divergence, such as KL divergence or Hellinger distance. Walker [Wal02] extends this approach by choosing the predictive distribution on the basis of a decision-theoretic criterion. The choice of divergence and the choice of the larger (nonparametric) model can be adhoc and are left up to the statistician.

Finally, we mention some recent work related to Bayesian model misspecification. Bühlmann and van de Geer [Bv15] investigate the robustness of asymptotic inference for misspecified linear models. Grünwald [Grü12] studies the use of power likelihood (i.e., adjusted likelihood) to improve the robustness to misspecification and proposes a method for choosing the power term. Miller and Dunson [MD15] propose the "coarsened posterior", obtained by conditioning on a KL neighborhood of the empirical distribution, rather than on the data directly, and show that the resulting posterior can be approximated by a power likelihood. Power likelihood approaches can obtain optimal convergence rates, but there is, as of yet, no clear motivation for the particular form of powered likelihood. Bissiri, Holmes, and Walker [BHW16] suggest a general framework for Bayesian inference and argue that a valid update of a prior belief distribution to a posterior can be made for parameters which are connected to observations through a loss function rather than the traditional likelihood function. Under model misspecification, the framework uses loss functions to connect information in the data to functionals of interest without building a more complex model.

4.2 Preliminaries

Let S, T be measurable spaces, and let $\mathcal{M}_1(S)$ denote the usual measurable space of probability measures on S. For $\mu \in \mathcal{M}_1(S)$ and $\tau : S \to T$ measurable, let $\tau_*\mu \in \mathcal{M}_1(T)$ denote the (pushforward) measure given by $(\tau_*\mu)(B) = \mu(\tau^{-1}[B]) = \mu\{s \in S : \tau(s) \in B\}$ for measurable $B \subseteq T$. Let κ be a probability kernel from S to T, i.e., a measurable map $\kappa : S \to \mathcal{M}_1(T)$. We will abuse notation and write $\kappa(B|s)$ for $(\kappa(s))(B)$ for all $s \in S$ and measurable $B \subseteq T$. Let $\mu \otimes \kappa$ denote the probability measure on $S \times T$ given by $(\mu \otimes \kappa)(A \times B) = \int_A \kappa(B|s)\mu(\mathrm{d}s)$, for all measurable $A \subseteq S$ and $B \subseteq T$. Finally, let $\mu^\tau : T \to \mathcal{M}_1(S)$ denote (some version of) the disintegration of μ along τ, i.e., μ^τ is a probability kernel from S to T satisfying

$$\mu\{s \in A : \tau(s) \in B\} = \int_B \mu^\tau(A|t)(\tau_*\mu)(\mathrm{d}t), \tag{4.1}$$

for all measurable $A \subseteq S$ and $B \subseteq T$. When $S = S_0 \times S_1$ is a product space, we will write μ_{S_i} for the pushforward of μ through the projection map $(s_0, s_1) \mapsto s_i$ and write $\mu_{S_{1-i}|S_i}$ for the disintegration of μ with respect to the same projection map. Let $\mu\kappa = (\mu \otimes \kappa)_T$ denote the marginal distribution induced on T, and note that $\mu \otimes \kappa = (\mu \otimes \kappa)_S \otimes (\mu \otimes \kappa)_{T|S}$.

4.2.1 Bayesian decision theory

In the classical definition of a statistical decision problem, one introduces a space of parameters Θ. The unknown loss of an action $a \in \mathcal{A}$ is defined to be a known function $L(\theta, a)$ of an unknown parameter $\theta \in \Theta$. The usual way to apply Bayesian decision theory is under the assumption that the model is well specified: one places a *prior* probability distribution on Θ and introduces a probability kernel on Θ to model the relationship between observations and the unknown parameter, so that they may serve as evidence about the losses we may incur after choosing an action. In particular, the model specifies a family of distributions, Q_θ, $\theta \in \Theta$, each defined on a common space, $\mathcal{X}$. Together with the probability kernel on Θ, a joint distribution on $\Theta \times \mathcal{X}$ is defined, which, given any observation, determines conditional expectations of the loss function for each potential action. In order to study Bayesian model misspecification, however, we need a more general setting in which there is no "parameter", which we now sketch. A detailed presentation is deferred to Section 4.4.1.

Let $(\bar{L}(\alpha))_{\alpha \in \mathcal{A}}$ be a $\mathbb{R}_+$-valued stochastic process, modeling unknown losses associated with available actions $\alpha \in \mathcal{A}$. Let ξ be a random element in a measurable space $\mathcal{X}$, representing evidence we have collected. Write μ for the distribution of ξ. Let $\mathcal{D}$ be the set of all maps $\delta : \mathcal{X} \to \mathcal{A}$ such that $\bar{L}(\delta(\xi))$ is a random variable, i.e., we have a belief about the loss we may incur upon responding to evidence ξ by taking action $\delta(\xi)$. Elements of $\mathcal{D}$ are called *decision procedures*, and the *Bayes risk* of $\delta \in \mathcal{D}$ is the expected loss, $\mathbb{E}\bar{L}(\delta(\xi))$, which may be infinite. A decision procedure δ^* is *Bayes* if $\mathbb{E}\bar{L}(\delta^*(\xi)) = \inf_{\delta \in \mathcal{D}} \mathbb{E}\bar{L}(\delta(\xi))$. Readers may be more familiar with Bayes (optimal) decision procedures being defined in terms of actions that minimize the posterior expected loss. We provide further details in Section 4.4.1. In particular, we shed more light on the relationship between these two definitions by revealing some structure in $\mathcal{D}$ using measure theory.

4.2.2 Some key results in Bayesian decision theory

We present several key results in Bayesian decision theory that we later reanalyze from a meta-Bayesian perspective.

Theorem 4.2.1. *Fix a σ-finite measure ν on $\mathcal{Y}$, let $\mathcal{A} = \{\nu' \in \mathcal{M}_1(\mathcal{Y}) : \nu' \ll \nu\}$ be the convex set of all distributions dominated by ν. Suppose Y is a random element in $\mathcal{Y}$ such that*

$$\bar{L}(\alpha) := -\log \frac{\mathrm{d}\alpha}{\mathrm{d}\nu}(Y), \quad \forall \alpha \in \mathcal{A}. \tag{4.2}$$

Assume that $P_{Y \mid \xi} \in \mathcal{A}$ for μ-almost all ξ, then the conditional distribution of Y given ξ is the unique Bayes optimal action, i.e., $\delta^*(\xi) = P_{Y \mid \xi}$ for μ-almost all ξ.

Let $m_i[\nu]$ denote the i-th moment of a distribution ν on $\mathbb{R}$, provided it exists. We will write $\mathrm{var}[\nu] := m_2[\nu] - (m_1[\nu])^2$ for the variance. For a distribution ν on $\mathbb{R}^d$, we will write $m[\nu]$ for the vector of coordinate-wise means and write $m_2[\nu]$ for the mean after the pushforward map $x \mapsto \|x\|_2^2$.

Theorem 4.2.2. *Let $\mathcal{Y} \subseteq \mathbb{R}^d$ and $\mathcal{A} = \mathrm{conv}(\mathcal{Y})$ be the convex hull of $\mathcal{Y}$. Suppose Y is a random element in $\mathcal{Y}$ such that*

$$\bar{L}(\alpha) := \|Y - \alpha\|_2^2, \quad \forall \alpha \in \mathcal{A}. \tag{4.3}$$

Assume that $m_2[P_{Y\mid\xi}] < \infty$ for μ-almost all ξ, then the mean of $P_{Y\mid\xi}$ is the unique Bayes optimal action, i.e., $\delta^*(\xi) = m[P_{Y\mid\xi}]$ for μ-almost all ξ.

4.3 Main results

4.3.1 Meta-Bayesian decision principle

The *Bayesian decision principle* says that, among a subclass $\mathcal{C} \subseteq \mathcal{D}$, of decision procedures, one should prefer $\delta_0 \in \mathcal{C}$ to $\delta \in \mathcal{C}$ if $\mathbb{E}\bar{L}(\delta_0(\xi)) \leq \mathbb{E}\bar{L}(\delta(\xi))$. In particular, if the infimum $\inf_{\delta\in\mathcal{C}} \mathbb{E}\bar{L}(\delta(\xi))$ is achieved by some element of $\mathcal{C}$, then we do not prefer using any procedure over this one.

Definition 4.3.1. A decision procedure δ^* is *meta-Bayes within* $\mathcal{C}$ if

$$\mathbb{E}\bar{L}(\delta^*(\xi)) = \inf_{\delta\in\mathcal{C}} \mathbb{E}\bar{L}(\delta(\xi)).$$

The meta-Bayesian decision principle can be used to study model misspecification. Under model misspecification, the subclass $\mathcal{C}$ includes all available decision procedures. The following example demonstrate the meta-Bayesian decision problem of choosing a prior for a misspecified model.

Example 4.3.2. (Meta-Bayesian decision problem of choosing a prior) We now consider a misspecified model Q and the *meta-Bayesian decision problem* of choosing a prior among a class $\mathcal{F}$. Every prior $\pi \in \mathcal{F}$ induces a joint belief on the losses $(\bar{L}(\alpha))_{\alpha\in\mathcal{A}}$ and evidence ξ. Thus, π determines which decision procedures are meta-Bayes within $\mathcal{F}$. Fix a Bayes decision procedure δ_π for each prior π. That is, for every $\pi \in \mathcal{F}$, let $\delta_\pi \in \mathcal{D}$ be Bayes under prior π and model Q. Having fixed a procedure for each prior, we see that *choosing* a prior π implicitly determines a procedure, and thus the loss $\bar{L}(\delta_\pi(\xi))$ that one will incur once observing ξ and taking the action $\delta_\pi(\xi)$. The *meta-Bayesian optimal prior* is then the prior that agrees with the Bayesian decision principle among $\{\delta_\pi : \pi \in \mathcal{F}\}$, i.e.,

$$\Pi = \arg\min_{\pi\in\mathcal{F}} \mathbb{E}\bar{L}(\delta_\pi(\xi)). \tag{4.4}$$

In the rest of the chapter, we will study Example 4.3.2 to demonstrate the meta-Bayesian principle under model misspecification. We will consider a particular setting for prediction tasks. More specifically, we assume that $\bar{L}(\alpha)$ can be written as $L(Y, \alpha)$ where Y is a random element in $\mathcal{Y}$ and $L(\cdot,\cdot)$ is a jointly measurable function on $\mathcal{Y}\times\mathcal{A}$. Writing the evidence ξ as X, the joint belief on the loss $\bar{L}(\alpha)$ and evidence ξ reduces to a joint belief on Y and X, which is a probability measure on $\mathcal{X}\times\mathcal{Y}$. More details are given in Section 4.3.2.

4.3.2 Model misspecification in prediction tasks

Let $\mathcal{X}$ and $\mathcal{Y}$ be measurable spaces, assumed to be Borel, let $\mathcal{A}$ be a space of actions, and let $L : \mathcal{Y}\times\mathcal{A} \to \mathbb{R}$ be a loss function, where $L(\cdot, a)$ is assumed to be measurable for every $a \in \mathcal{A}$. We are faced with the following task: We are shown the first element X in a pair (X, Y) of values in $\mathcal{X}\times\mathcal{Y}$. We are then asked to make a decision by choosing an action $A(X)$ in $\mathcal{A}$. Finally, we suffer a loss of $L(Y, A(X))$. Our goal is to choose actions so as to minimize our expected loss. We define the space of (nonrandomized) decision

procedures, denoted $\mathcal{D}$, to be the set of all functions $A : \mathcal{X} \to \mathcal{A}$ such that the map

$$(x, y) \mapsto L(y, A(x)) \tag{4.5}$$

is product measurable. Let P be a probability measure on $\mathcal{X} \times \mathcal{Y}$ representing our subjective uncertainty in the pair (X, Y). The Bayes risk of a decision procedure $A \in \mathcal{D}$ is defined to be

$$r(P, A) := \int_{\mathcal{X}\times\mathcal{Y}} L(y, A(x))\, P(\mathrm{d}(x, y)) \tag{4.6}$$

$$= \int_{\mathcal{X}} \Big\{ \int_{\mathcal{Y}} L(y, A(x))\, P_{\mathcal{Y}|\mathcal{X}}(\mathrm{d}y|x) \Big\} P^1(\mathrm{d}x). \tag{4.7}$$

A decision procedure $A \in \mathcal{D}$ is *Bayes optimal with respect to* P if the infimum

$$\inf_{A' \in \mathcal{D}} r(P, A') \tag{4.8}$$

is achieved by A. (Note that, in general, there may be zero, one, or more decision procedures satisfying this criterion.) Under some mild regularity conditions, we can describe Bayes optimal decision procedures directly: The Bayes optimal decision, conditioned on having observed some value x in $\mathcal{X}$, minimizes the conditional expected loss, i.e.,

$$A(x) \in \mathcal{A}[P, x] := \operatorname*{arg\,min}_{a \in \mathcal{A}} \int_{\mathcal{Y}} L(y, a)\, P_{\mathcal{Y}|\mathcal{X}}(\mathrm{d}y|x), \qquad \text{for } P^1\text{-almost all } x, \tag{4.9}$$

where $\operatorname{arg\,min}_{a\in\mathcal{A}} f(a) := \{a \in \mathcal{A} : \forall a' \in \mathcal{A}, f(a') \geq f(a)\}$ may be empty. We will assume that $\mathcal{A}[P, x]$ is a singleton set for P^1-almost all x and that Eq. (4.9) holds for some decision procedure $A \in \mathcal{D}$.[1] In this case, there is a unique (up to a P^1-null set) Bayes optimal decision procedure, which we will denote by A_P.

The following results are direct corollaries of Theorem 4.2.1 and Theorem 4.2.2.

Corollary 4.3.3 ([BS94, Proposition 3.14])**.** *Fix a σ-finite measure ν on $\mathcal{Y}$, let $\mathcal{A} = \{\nu' \in \mathcal{M}_1(\mathcal{Y}) : \nu' \ll \nu\}$ be the convex set of all distributions dominated by ν, and assume that $P_{\mathcal{Y}|\mathcal{X}}(x) \in \mathcal{A}$ for P^1-almost all x. Under the strictly convex* log loss *or* self-information loss*, $L(y, \xi) = -\log \frac{\mathrm{d}\xi}{\mathrm{d}\nu}(y)$, the posterior distribution is the unique Bayes optimal action, i.e., $A_P(x) = P_{\mathcal{Y}|\mathcal{X}}(x)$ for P^1-almost all x.*

Corollary 4.3.4 ([Sch96, Example 3.8])**.** *Let $\mathcal{Y} \subseteq \mathbb{R}^d$ and $\mathcal{A} = \mathrm{conv}(\mathcal{Y})$ be the convex hull of $\mathcal{Y}$, and assume that $m_2[P_{\mathcal{Y}|\mathcal{X}}(x)] < \infty$ for P^1-almost all x. Under the strictly convex* quadratic loss*, $L(y, z) = \|y - z\|_2^2$, the posterior predictive mean is the unique Bayes optimal action, i.e., $A_\pi(x) = m[P_{\mathcal{Y}|\mathcal{X}}(x)]$ for P^1-almost all x.*

Models and priors

In many practical applications of Bayesian analysis, one works with a belief P that takes the form of a mixture $P = \pi Q$, where $Q : \Theta \to \mathcal{M}_1(\mathcal{X} \times \mathcal{Y})$ is a probability kernel from a (Borel) space Θ of parameters and $\pi \in \mathcal{M}_1(\Theta)$. In general, there are many such decompositions. However, there is a unique

[1] If the action space $\mathcal{A}$ is a convex subset of a vector space and the loss $L(y, a)$ is a strictly convex function of a for every y, then there is no more than one Bayes optimal action, i.e., $A(x)$ is a singleton or empty set. See [Fer67, §2.5] for an abstract discussion of the existence of Bayes decision procedures.)

decomposition $P = P_{\mathcal{Y}}Q$ where $\mathcal{Y} = \Theta$ and $Q(\theta) = P_{\mathcal{X}|\mathcal{Y}}(\theta) \otimes \delta_\theta$. In standard nomenclature, Q is called a *model* and π a *prior* distribution.

Misspecified models

In practice, subjectivists can neither faithfully represent P nor leap directly to the optimal decision procedure A_P. Instead, pragmatism leads one to replace one's belief P by a mixture πQ over some tractable parametric family $Q = (Q_\theta)_{\theta\in\Theta}$ of distributions on $\mathcal{X} \times \mathcal{Y}$, measurably indexed by Θ, where the mixing weights are given by some distribution π on Θ.

We will refer to Q as *the model*, borrowing the terminology from classical statistics. This terminology may suggest that the model is somehow objective, or at least defines the problem at hand. This is often a mathematically productive stance, but in any real decision problem, π and Q are both *chosen* by the statistician and subject to (subjective) criticism insofar as $P \neq \pi Q$. Regardless, in many situations, substantially more effort is invested in the development of Q and so we proceed as if Q has been fixed. (Without much difficulty, the problem of model selection can be cast into this framework, although this is not the emphasis here.)

Fix a model Q. If, in fact, $P = \pi Q$ for some distribution π, i.e., if no approximation is necessary, then we are in the *well-specified* setting. In this case, in addition to the existence of random variables X and Y modeling our uncertainty in the pair, there is a random variable ϑ with distribution π such that Q_ϑ is (a regular version of) the conditional distribution of the pair (X, Y) given ϑ. The distribution π represents our subjective uncertainty on ϑ, i.e., π is our *prior* distribution, and Bayes optimal decisions can be expressed in terms of our posterior beliefs on ϑ given X. In particular,

$$P_{\mathcal{Y}|\mathcal{X}}(x) = \pi Q_{\mathcal{Y}|\mathcal{X}}(x) = \int_\Theta (Q_\theta)_{\mathcal{Y}|\mathcal{X}}(x)\, \pi(\mathrm{d}\theta|x), \qquad \text{for } P^1\text{-almost all } x, \tag{4.10}$$

where $\pi(\mathrm{d}\theta|x)$ is defined so as to make the second equality hold for $\pi Q_{\mathcal{X}}$-almost all x. In the well-specified setting, $\pi(\mathrm{d}\theta|X)$ is the conditional distribution of ϑ given X, and represents our posterior beliefs on ϑ given X.

The *misspecified* setting, where $P \neq \pi Q$ for any distribution π, is the typical one. In this case, there is no random variable ϑ such that Q_ϑ is the conditional distribution of (X, Y) given ϑ. As there is no such random variable ϑ, strictly speaking, there is no subjective prior distribution to speak of, and the use of subjectivist language in reference to π is inappropriate. At the very least, the subjective content of π is unclear and the subjective content of the "posterior" $\pi(\mathrm{d}\theta|x)$ is no clearer.

Irrespective of whether a model is misspecified, it is still possible that, for some distribution π, a decision procedure $A \in \mathcal{D}$ is Bayes optimal with respect to P if it is Bayes optimal with respect to πQ, in which case we will say that πQ *yields optimal decisions* (under loss L). In this case, using π as a "prior" in order to derive an action minimizing the "posterior" expected loss is arguably rational, at least for the particular loss function under consideration, even if the subjective content of π is unclear.

For example, if $P^1 \neq \pi Q_{\mathcal{X}}$ then the misspecification of Q does not necessarily contradict Eq. (4.10) holding for some distribution π and all versions of the disintegrations $P_{\mathcal{Y}|\mathcal{X}}$ and $\pi Q_{\mathcal{Y}|\mathcal{X}}$. In this case, call the model Q *conditionally well-specified*.[2] Then, for the appropriate choice of π, the conditional

[2]The reference to versions is a consequence of the fact that disintegrations are uniquely defined only up to null sets. When a canonical version of a disintegration exists, one might insist on agreement only for that version, although we do not explore this. This might be necessary if one were to model, e.g., a discrete variable with a continuous one.

distribution $\pi Q_{\mathcal{Y}|\mathcal{X}}(X)$ exactly matches the subjective conditional belief on Y given X, which implies that πQ yields optimal decisions for *every* loss function.

4.3.3 Meta-Bayesian decision problem for choosing priors

Let $\mathcal{F}$ denote a set of probability measures on Θ. For $\pi \in \mathcal{F}$, let A_π be the Bayes optimal decision procedure with respect to πQ, which we have assumed to be unique.[3] The *meta-Bayesian risk of* π, written $R(P, \pi)$, is the Bayes risk under P of performing Bayesian analysis under model Q with prior π, i.e.,

$$R(P, \pi) = r(P, A_\pi) = \int_{\mathcal{X}} \left\{ \int_{\mathcal{Y}} L(y, A_\pi(x)) P_{\mathcal{Y}|\mathcal{X}}(x) \right\} P^1(\mathrm{d}x). \tag{4.11}$$

We are interested in solutions to the *meta-Bayesian decision problem*, i.e., elements of the set

$$\Pi = \arg\min_{\pi \in \mathcal{F}} R(P, \pi), \tag{4.12}$$

which contains all π that achieve the minimum meta-Bayesian risk. Elements of Π will be called *optimal surrogate priors*.

We now specialize the meta-Bayesian decision problem to several classical loss functions: Recall that when $\mu \ll \nu$, the KL divergence from ν to μ is defined as $\mathrm{KL}\left(\mu || \nu\right) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\mu$.

Proposition 4.3.5. *Let $\nu \in \mathcal{M}(\mathcal{Y})$ be σ-finite, let $\mathcal{M}_1^F(\mathcal{Y})$ be the set of probability measures on $\mathcal{Y}$ that have finite differential entropy, let $\mathcal{A} = \{\mu \in \mathcal{M}_1^F(\mathcal{Y}) : \mu \ll \nu\}$, and assume that $\mathcal{A}$ contains $P_{\mathcal{Y}|\mathcal{X}}(x)$ and $\pi Q_{\mathcal{Y}|\mathcal{X}}(x)$ for P^1-almost all x and all $\pi \in \mathcal{F}$. Under log loss, as defined in Corollary 4.3.3, the meta-Bayesian decision problem is equivalent to minimizing the conditional relative entropy [Gra90], i.e.,*

$$\Pi = \arg\min_{\pi \in \mathcal{F}} \int \mathrm{KL}\left(P_{\mathcal{Y}|\mathcal{X}}(x) || \pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) P^1(\mathrm{d}x). \tag{4.13}$$

Proof. See Section 4.4.2. □

Remark 4.3.6. Gray [Gra90] gives a geometric interpretation of the conditional relative entropy, which can be used to translate the meta-Bayesian decision problem under log loss. In particular,

$$\int \mathrm{KL}\left(P_{\mathcal{Y}|\mathcal{X}}(x) || \pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) P^1(\mathrm{d}x) = \mathrm{KL}\left(P || P^1 \otimes \pi Q_{\mathcal{Y}|\mathcal{X}}\right). \tag{4.14}$$

Hence, the optimal prior π minimizes the KL divergence from $P^1 \otimes \pi Q_{\mathcal{Y}|\mathcal{X}}$ to P.

Furthermore, if there exists a parametric family $\{Q_\theta, \theta \in \Theta\}$ and a prior $\tilde{\pi}(\mathrm{d}\theta)$ such that $P^1 = \tilde{\pi} Q$ and $P_{\mathcal{Y}|\mathcal{X}}(x) = \tilde{\pi} Q_{\mathcal{Y}|\mathcal{X}}$, then minimizing the conditional relative entropy in Eq. (4.13) is equivalent to

$$\Pi = \arg\min_{\pi \in \mathcal{F}} \int_\Theta \left[\int_{\mathcal{X}} \mathrm{KL}\left(\pi_\theta Q_{\mathcal{Y}|\mathcal{X}}(x) || \pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) P_\theta(\mathrm{d}x) \right] \tilde{\pi}(\mathrm{d}\theta) \tag{4.15}$$

where π_θ is a degenerate distribution on $\{\theta\}$. Eq. (4.15) is first derived by Aitchison [Ait75] and is sometimes referred to as the Bayes risk under the KL loss [Kom96]. In [SDG06], the conditional relative

[3] Alternatively, we can fix some measurable selector $\pi \to A_\pi$ that identifies a Bayes optimal decision procedure for each prior choice. The meta-Bayesian risk is then that of A_π, and the meta-Bayesian decision problem is selecting decision procedures, not priors.

entropy $\int \mathrm{KL}\left(\tilde{\pi}Q_{\mathcal{Y}|\mathcal{X}}(x)||\pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) P^1(\mathrm{d}x)$ is defined as a function of $(\tilde{\pi}, \pi)$ and called the predictive relative entropy regret.

Similar forms to Eq. (4.15) has appeared in Bayesian predictive methods for model selection [PV17]. The so-called reference model approach is to construct a full encompassing reference model, which is believed to best describe our knowledge about the future observations, and perform the utility estimation almost as if it was the true data generating distribution. Then one idea for model selection is first to project the information in the posterior of the reference model onto the candidate models and then select the candidate model whose predictive distribution is closest to the reference model. See, for example, [VO12; GR98] for more details.

Other problems that have similar forms to Eq. (4.15) have appeared in existing literature. This includes some work on nonsubjective priors [Gho11] and minimax shrinkage predictive algorithms [GLX12]. The main difference between those work and our work is threefold. First, most existing work assumes the model is well-specified, in this work, however, we study model misspecification; second, although in some work the objective has the form of KL divergence, but it is usually not the KL divergence of predictive distributions; finally, nonsubjective priors have been widely studied through a minimax approach, which is different from our work since we study subjective priors by directly minimizing the conditional relative entropy. In the following, we give a literature review of such existing work in both areas of nonsubjective priors and shrinkage prediction.

There is a subclass of nonsubjective priors named "divergence priors" due to Ghosh [Gho11]. The loss function defined using the KL divergence is studied, such as in the reference priors [Ber79; BBS09]. Another example is the maximum likelihood prior by Hartigan [Har98], in which a truncated KL loss is defined. However, the aforementioned work is not based on KL divergence of predictive distributions. The most relevant work in nonsubjective priors is the work by Sweeting, Datta, and Ghosh [SDG06]. They suggest to use posterior relative entropy regret for deriving nonsubjective prior. It is argued that the reference prior criterion [Ber79] based on scoring the prior predictive distribution is less relevant than the posterior predictive distribution. This is because we are not so much interested in predicting the data already observed as future data yet to be observed. More details can be found in [SDG06] and also the discussion of the paper [Gho11] by Sweeting [Swe11].

Without taking expectation over θ in Eq. (4.15), the expected KL divergence over $x \in \mathcal{X}$ for fixed θ is defined as the KL loss in the development of shrinkage estimators for multivariate normal predictive densities [GLX06; BGX08]. This work can be seen as parallel developments of the classical results on minimax shrinkage estimators under quadratic loss in [Bro71; BH82; Ste81]. A review of the work in this area can be found in [GLX12]. For models that satisfies some differential geometric conditions, asymptotic analysis when the number of data goes to infinity is studied [Kom96; Kom06] using the tool of information geometry [AN00]. For example, Komaki [Kom06] shows that there exists a shrinkage prior that asymptotically gives smaller KL loss than the Jeffreys prior, for any given θ. However, there is very little work that studies model misspecification. Perhaps the most relevant work is by Fushiki [Fus05]. Fushiki [Fus05] compares the Bayesian predictive distribution with the Bootstrap prediction obtained by applying Breiman's "bagging" method to a plug-in prediction. The result of [Fus05] is for the asymptotic cases where the number of observed data goes to infinity. It is worth to further investigate the work in [Fus05] and its relations to both [Mül13] and [RT03]. It is promising that the rationality of the methods in [Mül13] and [RT03] can be analyzed in the framework of meta-Bayesian analysis.

Proposition 4.3.7. *Under quadratic loss, the meta-Bayesian decision problem is equivalent to minimizing*

the expected quadratic distance between the true and model posterior means, i.e.,

$$\Pi = \arg\min_{\pi \in \mathcal{F}} \int \|m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)] - m[P_{\mathcal{Y}|\mathcal{X}}(x)]\|_2^2 \, P^1(\mathrm{d}x). \tag{4.16}$$

Proof. See Section 4.4.2. □

Optimal surrogate priors depend on the loss

From Propositions 4.3.5 and 4.3.7, we can see that the meta-Bayesian decision problem takes different forms under different loss functions. In this section, we investigate the set of optimal surrogate priors under various hypotheses and show that the set of optimal surrogate priors depends, in general, on the loss function.

We begin with the following result, which characterizes the set of optimal priors in the case where the model can yield optimal decisions for some prior:

Proposition 4.3.8. *Let* $\tilde{\Pi} = \{\pi \in \mathcal{F} : A_\pi(x) = A_P(x) \text{ for } P^1\text{-a.a. } x\}$. *If* $\tilde{\Pi} \neq \emptyset$, *then* $\Pi = \tilde{\Pi}$.

Proof. See Section 4.4.2. □

The following corollary implies that meta-Bayesian analysis reduces to ordinary Bayesian analysis in the well-specified case. In other words, one's own subjective prior is optimal, provided it is in the class $\mathcal{F}$ of priors under consideration.

Corollary 4.3.9 (Self-consistency)**.** *If there exists* $\pi \in \mathcal{F}$ *such that* $\pi Q_{\mathcal{Y}|\mathcal{X}} = P_{\mathcal{Y}|\mathcal{X}}$ *on a* P^1*-measure one set, then* $\pi \in \Pi$. *In particular, if* $P = \pi Q$ *for some* $\pi \in \mathcal{F}$, *then* $\pi \in \Pi$.

Proof. If $\pi Q_{\mathcal{Y}|\mathcal{X}} = P_{\mathcal{Y}|\mathcal{X}}$ on a P^1-measure one set, then by the definitions of $A_\pi(x)$ and $A_P(x)$, we have $A_\pi(x) = A_P(x)$ for P^1-a.a. x. Therefore, $\pi \in \Pi$ from Proposition 4.3.8. □

The following corollaries demonstrate that Π can depend on the choice of the loss function:

Corollary 4.3.10. *Under log loss,*

$$\Pi = \{\pi \in \mathcal{F} : P_{\mathcal{Y}|\mathcal{X}}(x) = \pi Q_{\mathcal{Y}|\mathcal{X}}(x) \text{ for } P^1\text{-a.a. } x\}$$

if the r.h.s. is not empty.

Corollary 4.3.11. *Under quadratic loss,*

$$\Pi = \{\pi \in \mathcal{F} : m[P_{\mathcal{Y}|\mathcal{X}}(x)] = m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)] \text{ for } P^1\text{-a.a. } x\}$$

if the r.h.s. is not empty.

From Corollaries 4.3.10 and 4.3.11 it follows that, if πQ yields optimal decisions under log loss, then πQ yields optimal decisions under quadratic loss, provided the first moments exist. However, the reverse implication does not hold in general, and so the set of optimal surrogate priors can, in general, depend on the loss.

The following example is of a model that yields optimal decisions under squared loss, but does not yield optimal decisions under log loss:

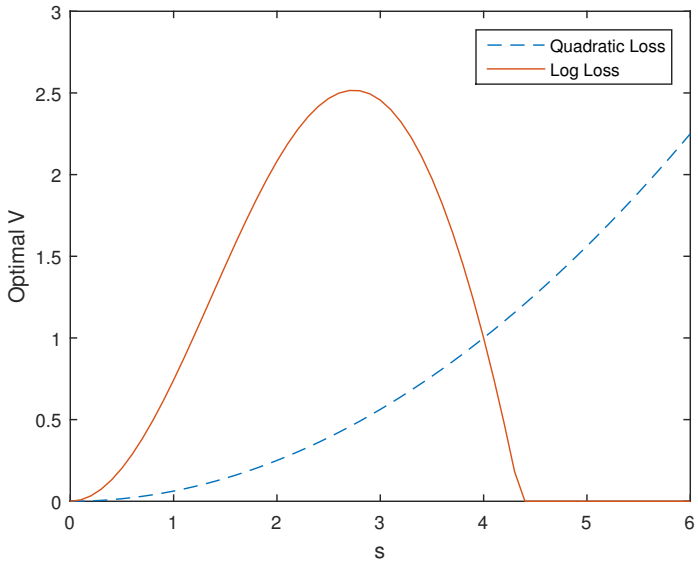

Figure 4.1: Optimal zero-mean normal prior for simple normal model with misspecified variance when true standard deviation is set to $r = 4$.

Example 4.3.12. Let the true belief be $P_\theta = \mathcal{N}(\theta, r^2)$ with prior belief $\tilde{\pi} = \mathcal{N}(0, 1)$. Suppose the model $Q_\theta = \mathcal{N}(\theta, s^2)$ where $s^2 > 0$ may not equal r^2. We consider the optimal prior on Q_θ that has the form $\pi = \mathcal{N}(0, V)$ with the only parameter V. The predictive distributions of Y given X under P and πQ are, respectively,

$$\mathcal{N}\Big(\frac{x}{1+r^2}, r^2 + \frac{r^2}{1+r^2}\Big) \quad \text{and} \quad \mathcal{N}\Big(\frac{x}{1+s^2/V}, s^2 + \frac{s^2}{1+s^2/V}\Big). \tag{4.17}$$

The optimal value of V for different values of s when $r = 4$ is shown in Fig. 4.1. According to the figure, one can see that the optimal zero-mean normal prior depends on the loss function. The only case when the optimal values of V coincide is when the model is well-specified ($s = r = 4$), in which case the meta-Bayesian analysis reduces to traditional Bayesian analysis. Note that under the quadratic loss, the optimal $V = \frac{s^2}{r^2}$, which matches the predictive mean. The optimal V under log loss is less intuitive. For example, the optimal V is zero when $s^2 \geq r^2 + 3$. In this case, the predictive distribution for Y reduces to $\mathcal{N}(0, s^2)$, which does not depend on x. This scenario can be intuitively explained as follows. Because minimizing the log loss results in minimizing the conditional relative entropy, when s^2 is much larger than r^2, the optimal prior is chosen to match the predictive variance, ignoring the predictive mean.

Optimal surrogate priors are not necessarily true beliefs

Next, we consider the case that the parameter of the misspecified model has certain realistic meaning such that one actually has a "true belief" on it. We demonstrate that this "true belief" may not be the optimal surrogate prior for the misspecified model.

Consider a $\{0, 1\}$-valued sequence $\{X_i\}$. As our true belief, consider $\{X_{2i-1}, i = 1, 2, \dots\}$ are i.i.d. Bernoulli random variables with parameter θ, and $X_{2i-1} = X_{2i}, \forall i$. Suppose our true belief on θ follows a uniform distribution on $[0, 1]$. Then we show that this uniform prior may not be the optimal surrogate

prior when the misspecified model is the i.i.d. Bernoulli model.

Example 4.3.13. (Prediction using i.i.d. Bernoulli model) Suppose we aim to predict $Y = X_2$ after observing $X = X_1$ using the misspecified i.i.d. Bernoulli model with parameter θ. Under the log loss, the optimal surrogate prior is given by Eq. (4.13). Clearly, a prior that concentrates on $\{0, 1\}$ is a solution of Eq. (4.13) since the prediction using such a prior is always $Y = X$, which is optimal. The uniform prior is clearly not the optimal surrogate prior, even though it is the true belief on θ.

Example 4.3.14. (Inference using i.i.d. Bernoulli model) Suppose we aim to construct an inference for $Y = \theta$ after observing $X = (X_1, X_2)$ using the misspecified i.i.d. Bernoulli model with parameter θ. Under the quadratic loss, the optimal surrogate prior is given by Eq. (4.16). Under a Beta prior $\pi \sim \mathrm{Beta}(\alpha, \beta)$,

$$\int \|m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)] - m[P_{\mathcal{Y}|\mathcal{X}}(x)]\|_2^2 \, P^1(\mathrm{d}x) = \mathbb{E}\left(\frac{\alpha + 2X_1}{\alpha + \beta + 2} - \frac{1 + X_1}{3}\right)^2, \tag{4.18}$$

where the expectation is taken over $X_1 \sim \mathrm{Beta}(1, 1)$. Clearly, the uniform prior $\alpha = \beta = 1$ is not the optimal surrogate prior since

$$\mathbb{E}\left(\frac{1 + 2X_1}{1 + 1 + 2} - \frac{1 + X_1}{3}\right)^2 = \mathbb{E}\left(\frac{2X_1 - 1}{12}\right)^2 \doteq 0.0231, \tag{4.19}$$

is dominated by the optimal surrogate prior $\mathrm{Beta}(2, 2)$, which leads to

$$\mathbb{E}\left(\frac{2 + 2X_1}{2 + 2 + 2} - \frac{1 + X_1}{3}\right)^2 = 0. \tag{4.20}$$

4.3.4 Asymptotically optimal priors for i.i.d. models and i.i.d. beliefs

Existing results on the asymptotic behavior of posterior distributions under misspecification can suggest the form of asymptotically optimal surrogate priors. In particular, Berk [Ber66], building on the work of LeCam [LeC53], studied dominated i.i.d. models under misspecification and showed that, under some regularity conditions, the posterior distribution of the parameter converges weakly to the set (called the asymptotic carrier) of parameters of the model that minimize the Kullback–Leibler divergence with respect to the data-generating distribution (which is assumed to be i.i.d. as well). The asymptotic carrier may, in general, contain more than one point, and may depend on both the model and the data-generating distribution. When the model is well-specified and identifiable, the asymptotic carrier contains only one point, which indexes the data-generating distribution. In general, however, there need be no convergence (in any sense) of the posterior distribution to a limiting distribution over the asymptotic carrier. The conditions under which a sequence of posterior distributions converges weakly to a degenerate distribution were further studied by Berk [Ber70]. Later, Bunke and Milhaud [BM98] strengthened the convergence beyond a.s. weak convergence under additional hypotheses.

In the well-specified and identifiable case, our prior distribution can be equivalently characterized as our belief on the asymptotic location of the posterior distribution. As a heuristic in the misspecified setting, our surrogate prior can be taken to be our personal belief as to the asymptotic location of the model's posterior distribution, which depends on a surrogate prior only through its support under regularity conditions. Our result shows that this heuristic is indeed meta-Bayesian optimal asymptotically.

We begin with the classic result due to Berk [Ber66].

Lemma 4.3.15. *Let $\mathcal{X}$ and Θ be Borel subsets of complete separable metric spaces. Let Q_θ, for $\theta \in \Theta$, be a measurable family of probability measures on $\mathcal{X}$, dominated by a σ-finite measure μ with conditional density $g(\cdot\,|\,\cdot) : \mathcal{X} \times \Theta \to \mathbb{R}_+$. Let π be a prior distribution on the Borel subsets of Θ, and assume π has full support.*

Let F be a probability measure on $\mathcal{X}$ which admits a density g^* with respect to μ, and let F^∞ denote the infinite product measure on $\mathcal{X}^\infty$. Assume:

1. *for F-almost all x, $g(x\,|\,\cdot)$ is continuous in θ;*
2. *for all $\theta \in \Theta$, $F\{x : g(x\,|\,\theta) > 0\} = 1$;*
3. *for all $\theta \in \Theta$, there is an open neighborhood U of θ such that*

$$\int \sup_{\theta \in U} \left\{ \log \frac{g(x\,|\,\theta)}{g^*(x)} \right\} F(\mathrm{d}x) < \infty; \text{ and} \tag{4.21}$$

4. *there is a positive integer k such that for every $r \in \mathbb{R}$, there is a co-compact subset $D \subseteq \Theta$ such that*

$$\int_{\mathcal{X}^k} \sup_{\theta \in D} \left\{ \frac{1}{k} \sum_{i=1}^{k} \log \left[\frac{g(x_i\,|\,\theta)}{g^*(x_i)} \right] \right\} F(\mathrm{d}x_1) \cdots F(\mathrm{d}x_k) \leq r. \tag{4.22}$$

Define $A_0 = \arg\min_{\theta \in \Theta} \mathrm{KL}\,(F||Q_\theta)$ to be the set of all $\theta \in \Theta$ that achieves the minimum. For every open set $U \supseteq A_0$,

$$\lim_{k \to \infty} \frac{\int_U \prod_{i=1}^k g(x_i\,|\,\theta)\pi(d\theta)}{\int_\Theta \prod_{i=1}^k g(x_i\,|\,\theta)\pi(d\theta)} = 1, \quad F^\infty\text{-a.a. } x. \tag{4.23}$$

Proof. See [Ber66, Main theorem]. □

Suppose the true belief can be written as a mixture $P_\infty = \tilde{\nu}\tilde{P}$, $= \int \tilde{P}_\psi \tilde{\nu}(\mathrm{d}\psi)$, where $\tilde{P}_\psi$ is i.i.d. For each $\psi \in \Psi$, there is a unique parameter $\phi(\psi) \in \Theta$ such that $Q_{\phi(\psi)}$ is the distribution minimizing the KL divergence with $\tilde{P}_\psi$. Viewing ϕ as a many-to-one mapping from Ψ to Θ, let $\tilde{\pi} = \tilde{\nu} \circ \phi^{-1}$ be the pushforward of $\tilde{\nu}$ through ϕ, and let $\tilde{\nu}^\phi(\mathrm{d}\psi)$ be the disintegration of $\tilde{\nu}$ along the map ϕ. Defining $P_\theta = P(\theta, \cdot) = \int \tilde{P}_\psi\, \tilde{\nu}^\phi(\mathrm{d}\psi|\theta)$, we can write the true belief as $P_\infty = \int_\Theta P_\theta \tilde{\pi}(\mathrm{d}\theta) = \int_\Theta P(\theta, \cdot)\tilde{\pi}(\mathrm{d}\theta) = \tilde{\pi}P$. Note that nested models are special cases of our formulation when $\Theta \subset \Psi$.

In this section, we will write $P_\theta^{(k)}$ to denote the restriction of P_θ to $\mathcal{X}^k$ obtained by marginalizing out all but the first k elements of the sequence. We will also define the kernel product Q^k by $Q^k(\theta, \cdot) = (Q_\theta)^k$ for every $\theta \in \Theta$. Next, we show our main result on the asymptotically optimal prior when $n = 0$ and $k \to \infty$.

Theorem 4.3.16. *Let $\mathcal{X}$, Θ, Q, μ and $g(\cdot\,|\,\cdot)$ be as in Lemma 4.3.15. Let P_θ, for $\theta \in \Theta$, be a measurable family of probability measures on $\mathcal{X}$ admitting a conditional density $g^*(\cdot\,|\,\cdot) : \mathcal{X} \times \Theta \to \mathbb{R}_+$ w.r.t. μ, such that P_θ and $g^*(\cdot\,|\theta)$ satisfy the conditions on F for every θ.*

Let $\tilde{\pi}$ be a probability measure on Θ that admits a density $f_{\tilde{\pi}}$ w.r.t. μ and has full support. Assume

5. *For every $\theta \in \Theta$, the point θ is the unique point in Θ achieving the infimum $\inf_{\theta'' \in \Theta} \mathrm{KL}\,(Q_{\theta''}||P_\theta)$.*

6. *Let $\mathcal{F}$ be the set of all priors that (a) admit a density f_π; (b) have full support; (c) $\frac{f_\pi(\theta)}{f_{\tilde{\pi}}(\theta)}$ is continuous w.r.t. θ and has a continuous extension to $\bar{\Theta}$, the closure of Θ; (d) $\sup_{\theta \in U}\{\frac{f_\pi(\theta)}{f_{\tilde{\pi}}(\theta)}\} < \infty$ for all open sets $U \subseteq \Theta$.*

Then $\tilde{\pi}$ is the asymptotically optimal prior in the sense that as $k \to \infty$,

$$\mathrm{KL}\left(\tilde{\pi}P^{(k)}||\tilde{\pi}Q^k\right) - \inf_{\pi \in \mathcal{F}} \mathrm{KL}\left(\tilde{\pi}P^{(k)}||\pi Q^k\right) \to 0. \tag{4.24}$$

Proof. See Section 4.4.2. □

If we assume that all the conditions in Proposition 4.3.5 hold so that $R(\pi P^{(k)}, \pi) - R(\pi P^{(k)}, \pi') = \mathrm{KL}\left(\pi P^{(k)}||\pi Q^k\right) - \mathrm{KL}\left(\pi P^{(k)}||\pi' Q^k\right)$, for every π, π', then this implies that $\tilde{\pi}$ is asymptotically optimal.

Other results on misspecification may yield results in meta-Bayesian analysis: In recent work, Kleijn and van der Vaart [Kv06] extended the results of [Ber66; Ber70; BM98] to characterize the asymptotic behavior and rate of convergence of the posterior distribution in the setting of infinite-dimensional nonparametric models under misspecification. Recently, Kleijn and van der Vaart [Kv12] showed that, the posterior distribution under misspecified parametric models is approximately normal as in the well-specified case, however, the covariance structure under misspecification results in credible regions failing to have the desired frequentist coverage. Under squared loss, we would conjecture that optimal surrogate priors would be those that, on average, tend to correct the covariance structure.

4.3.5 Optimal priors when the i.i.d. Bernoulli model is well-specified

By Corollary 4.3.9, when a model is well-specified, the set of optimal priors contains the true prior. However, there will, in general, be additional surrogate priors. By characterizing the set of optimal surrogate priors, we can gain insight into what aspects of our prior can possibly affect our inferences, which can reduce the complexity of the problem of eliciting prior beliefs for the task at hand.

In this section, we investigate sufficient conditions for the optimality of surrogate priors in the well-specified i.i.d. Bernoulli setting (Section 4.3.5). Fix $n, k \in \mathbb{N}$, and let $\mathbb{B} = \{0, 1\}$, $\mathcal{X} = \mathbb{B}^n$, and $\mathcal{Y} = \mathbb{B}^k$. Recall that n is the number of Bernoulli observations we will observe before being asked to make a prediction, and k is the number of Bernoulli observations we will be asked to predict. Let $Q = (Q_\theta)_{\theta \in \Theta}$ be the family of all (Bernoulli) product measures on $\mathcal{X} \times \mathcal{Y}$ indexed by their common mean in $\Theta = [0, 1]$, and assume that P is some mixture $\tilde{\pi}Q$ of (Bernoulli) product measures on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F} \subseteq \mathcal{M}_1(\Theta)$ be the set of priors under consideration.

Sufficient condition for optimal surrogate priors

Recall that $m_i[\mu]$ denotes the i-th moment of μ. We have the following results:

Proposition 4.3.17. *Under log loss, if the r.h.s. is not empty, then*

$$\Pi = \{\pi \in \mathcal{F} : m_i[\pi] = m_i[\tilde{\pi}], \forall i = 1, \ldots, n + k\} \tag{4.25}$$

Proof. See Section 4.4.2. □

That is, if feasible, the optimal surrogate priors are exactly those priors that match the first $n + k$ moments of the prior belief $\tilde{\pi}$.

Proposition 4.3.18. *Under quadratic loss, if the r.h.s. is not empty, then*

$$\Pi = \{\pi \in \mathcal{F} : m_i[\pi] = m_i[\tilde{\pi}], \forall i = 1, \ldots, n+1\} \tag{4.26}$$

Proof. See Section 4.4.2. □

That is, under quadratic loss, if feasible, a surrogate prior is optimal if it matches the first $n+1$ moments of the prior belief. Note that Π does not depend on k under the quadratic loss. If the Π in Proposition 4.3.17 is not empty, then the Π under the log loss is a subset of the Π under the quadratic loss.

4.3.6 The i.i.d. Bernoulli model when exchangeability does not hold

In this section, we investigate optimal surrogate priors for the i.i.d. Bernoulli model when the underlying exchangeability assumption does not hold under the subjective belief P. We will take $n = 0$, $k \geq 1$, and study log loss. It follows that the meta-Bayesian decision problem must reason about sequential forecasting: at each stage, we must produce a probability for the next Bernoulli observation and the loss is the sum of our log losses across stages. Our belief does not assume that the observations are conditionally i.i.d., but the model does, and so the model will not, in general, yield optimal decisions.

Let $Q = (Q_\theta)_{\theta \in [0,1]}$ denote the family of all Bernoulli measures on $\mathbb{B}$, indexed by their means, and write Q_θ^k and Q_θ^∞ for the k-product and infinite product measures, respectively, i.e., the distributions of length-k and infinite i.i.d. Bernoulli sequence, each element having mean θ. Let $\vartheta_k(x) := \frac{1}{k}\sum_j x_j$ for $x \in \mathbb{B}^\infty$ and $k \in \mathbb{N}$. By the law of large numbers, $\vartheta_k(x) \to \theta$ as $k \to \infty$ for Q_θ^∞-almost all x. In other words, the limiting relative frequency $\vartheta = \lim \vartheta_k$ exists almost surely and coincides with the parameter θ almost surely.

Consider an arbitrary distribution P_∞ on $\mathbb{B}^\infty$, subject to the condition that $\vartheta(x)$ exists for P_∞-almost all x. Note that the existence of the limiting relative frequency is implied by exchangeability, but does not imply exchangeability. Instead, we will say that P_∞ merely *believes in a limiting relative frequency*. In this case, we define $\tilde{\pi}$ to be the P_∞-distribution of ϑ, i.e., the pushforward of P_∞ through the map ϑ, and define the map $\theta \mapsto P_\theta^{(\infty)}$ to be the disintegration of P_∞ with respect to ϑ, which then satisfies $\tilde{\pi}P^{(\infty)} = P_\infty$. We assume that, uniformly over $\theta \in [0,1]$, $\vartheta_k(x) \to \theta$ as $k \to \infty$ for P_θ^∞-almost all x. For $k \in \mathbb{N}$, let $P^{(k)}$ be the restriction of $P^{(\infty)}$ to $\mathbb{B}^k$ obtained by marginalizing out all but the first k elements of the sequence. Let $\mathcal{M}_1^c[0,1]$ be the set of all absolutely continuous probability measures π on $[0,1]$ whose density functions are Lipschitz continuous.

Theorem 4.3.19. *If $\tilde{\pi} \in \mathcal{M}_1^c[0,1]$, then $\tilde{\pi}$ is an asymptotically optimal prior with respect to $\mathcal{M}_1^c[0,1]$, i.e., as $k \to \infty$,*

$$R(\tilde{\pi}P^{(k)}, \tilde{\pi}) - \inf_{\pi \in \mathcal{M}_1^c[0,1]} R(\tilde{\pi}P^{(k)}, \pi) \to 0. \tag{4.27}$$

Proof. See Section 4.4.2. □

Note that when $n \neq 0$ and/or k is small, $\tilde{\pi}$ may be far from optimal. The following example illustrates this fact.

Example 4.3.20. Consider a $\{0,1\}$-valued discrete-time Markov chain with transition matrix

$$\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}. \tag{4.28}$$

Then the limiting relative frequency of 1's is $\theta := \frac{p}{p+q}$ and the second eigenvalue of the transition matrix is $1-p-q$, and these two parameters determine p and q. (Note that the Markov chain is i.i.d. if and only if the second eigenvalue is zero, i.e., if and only if $p+q=1$.) As our true belief P_∞, consider a mixture of such Markov processes, where $2\theta \sim \text{Beta}(a,b)$ and $p+q \sim \text{Beta}(c,d)$, independently. Note that the independence assumption is valid since $\theta \in (0,1/2)$ implies $p = \theta(p+q) \in (0,1/2)$ and $q = (1-\theta)(p+q) \in (0,1/2)$. In this case, the spectral gap $\nu := 1-|1-p-q|$ reduces to $\nu = p+q \sim \text{Beta}(c,d)$. The meta-Bayesian risk of using an i.i.d. model with prior $\text{Beta}(a,b)$ can be written in closed form as a function of the parameters (a,b,c,d). We will exploit this to understand the role of the spectral gap.

For simplicity, take $a=b>0$, hence $\mathbb{E}\theta = \frac{1}{4}$. A typical chain could mix very slowly when the expected spectral gap $\mathbb{E}\nu = \frac{c}{c+d}$ is small and an i.i.d. sequence can be seen as the special case where $\mathbb{E}\nu = 1$. Next, we consider the case of $n=0$ and $k=2$, where the true belief $\tilde{\pi}P^{(2)}$ is

$$\tilde{\pi}P^{(2)}(\{11\}) = \frac{1+a}{8(1+2a)}\mathbb{E}\nu + \frac{1}{4a}(1-\mathbb{E}\nu) \tag{4.29}$$

$$\tilde{\pi}P^{(2)}(\{00\}) = \frac{3}{4} - \frac{1+3a}{8(1+2a)}\mathbb{E}\nu \tag{4.30}$$

$$\tilde{\pi}P^{(2)}(\{10\}) = \tilde{\pi}P^{(2)}(\{01\}) = \frac{1+3a}{8(1+2a)}\mathbb{E}\nu. \tag{4.31}$$

For example, if the true belief is $a=c=1$ and $d=1000$, the Markov chain is very "sticky" (i.e., $\mathbb{E}\nu = \frac{1}{1001}$ is small). Under the i.i.d. model with a $\text{Beta}(1,1)$ prior, the probability of $\{01,10\}$ is higher and of $\{11,00\}$ is lower. Indeed, because $\frac{1+a}{8(1+2a)}$ is decreasing and $\frac{1+3a}{8(1+2a)}$ is increasing with a, it can be verified that for all $x \in \{0,1\}^2$, we have

$$\frac{\mathrm{d}}{\mathrm{d}a'}\left[\tilde{\pi}P^{(2)}(x)\log\frac{\tilde{\pi}P^{(2)}(x)}{\pi Q^{(2)}(x)}\right]\Bigg|_{a'=1} > 0, \tag{4.32}$$

where π denotes the beta prior $\text{Beta}(a',a')$. This implies

$$\frac{\mathrm{d}}{\mathrm{d}a'}R(P,\pi)\Big|_{a'=1} = \frac{\mathrm{d}}{\mathrm{d}a'}\,\text{KL}\left(\tilde{\pi}P^{(2)}||\pi Q^{(2)}\right)\Big|_{a'=1} > 0. \tag{4.33}$$

Therefore, choosing a $\text{Beta}(a',a')$ prior for some $a'<1$ leads to a lower meta-Bayesian risk. Indeed, $a'=0.001$ is preferred to $a'=1$, because $R(P,\pi) = \text{KL}\left(\tilde{\pi}P^{(2)}||\pi Q^{(2)}\right) \doteq 0.444$ bits while $R(P,\tilde{\pi}) = \text{KL}\left(\tilde{\pi}P^{(2)}||\tilde{\pi}Q^{(2)}\right) \doteq 0.666$ bits. The situation is similar when $n=1$ and $k=1$: in this case, $R(P,\pi) = 1.2543$ bits, while $R(P,\tilde{\pi}) = 1.9363$ bits.

4.4 Appendices

4.4.1 Bayesian decision theory in general settings

In order to shed more light on the Bayes relation, we begin by revealing some structure in $\mathcal{D}$: For every $\alpha \in \mathcal{A}$, let $\check{\alpha}$ denote the constant map from $\mathcal{X}$ to $\mathcal{A}$, taking every point in $\mathcal{X}$ to α. For every measurable partition $B_1, B_2, \cdots \subseteq \mathcal{X}$ of $\mathcal{X}$ and sequence $\delta_1, \delta_2, \cdots \in \mathcal{D}$, let $[\delta_1 : B_1, \delta_2 : B_2, \dots]$ denote the map from $\mathcal{X}$ to $\mathcal{A}$ sending $x \in B_n$ to $\delta_n(x)$. The we have the following result (the proof is given in Section 4.4.2).

Lemma 4.4.1. *For every $\alpha \in \mathcal{A}$, $\check{\alpha} \in \mathcal{D}$. For every measurable partition $B_1, B_2, \cdots \subseteq \mathcal{X}$ of $\mathcal{X}$ and sequence $\delta_1, \delta_2, \cdots \in \mathcal{D}$, $[\delta_1 : B_1, \delta_2 : B_2, \dots] \in \mathcal{D}$.*

For a random variable V, write $\mathbb{E}[V; \xi \in B]$ for $\mathbb{E}(1_{\{\xi \in B\}} V)$ and let $\sigma(V)$ denote the σ-algebra generated by V. Let $\delta \in \mathcal{D}$ and let $\mathbb{E}^{\xi} \bar{L}(\delta(\xi))$ denote the conditional expectation of $\bar{L}(\delta(\xi))$ given ξ, i.e., $\mathbb{E}^{\xi} \bar{L}(\delta(\xi))$ is a $\sigma(\xi)$-measurable random variable and, for all measurable $B \subseteq \mathcal{X}$, $\mathbb{E}[\mathbb{E}^{\xi} \bar{L}(\delta(\xi)); \xi \in B] = \mathbb{E}[\bar{L}(\delta(\xi)); \xi \in B]$. In particular, we have the chain rule, $\mathbb{E}\mathbb{E}^{\xi} \bar{L}(\delta(\xi)) = \mathbb{E}\bar{L}(\delta(\xi))$. By the $\sigma(\xi)$-measurability, there is a measurable function $\ell_\delta : \mathcal{X} \to \mathbb{R}_+$ such that $\mathbb{E}^{\xi} \bar{L}(\delta(\xi)) = \ell_\delta(\xi)$. We call ℓ_δ the posterior risk function.

We now describe the Bayes relation in terms of lattice infimum: Consider a measured space $(\mathcal{X}, \nu)$ and the quotient of the space of real-valued measurable functions on $\mathcal{X}$ under the equivalence relation $f \sim g$ iff $f = g$ ν-a.e. Given a class $\mathcal{F}$ of real-valued measurable functions on $(\mathcal{X}, \nu)$, recall that the *lattice infimum of* $\mathcal{F}$ is (the equivalence class of) a measurable function $\bigwedge \mathcal{F} : \mathcal{X} \to \mathbb{R}_+$ that satisfies

$$\forall f \in \mathcal{F},\ \bigwedge \mathcal{F} \leq f \ \ \nu\text{-a.e.} \tag{4.34}$$

$$\forall \text{ measurable } g,\ \left(\forall f \in \mathcal{F},\ g \leq f \ \ \nu\text{-a.e.}\right) \implies \left(g \leq \bigwedge \mathcal{F} \ \ \nu\text{-a.e.}\right). \tag{4.35}$$

Note that, if $\mathcal{F}$ is countable, then $\inf \mathcal{F}$ is measurable and $\inf \mathcal{F} = \bigwedge \mathcal{F}$ ν-a.e. In general, $\inf \mathcal{F}$ may not be measurable. If $\mathcal{X}$ is a Borel subset of $\mathbb{R}^n$, and $\mathcal{F}$ is a class of real-valued measurable functions on $\mathcal{X}$, then the lattice infimum of $\mathcal{F}$ exists and is equivalent to the pointwise infimum of some countable subset $\mathcal{F}' \subseteq \mathcal{F}$ due to Hajłasz and Malý [HM02, Lemma 2.6]. Then we have the following equivalent statements to being Bayes. The proof is given in Section 4.4.2.

Lemma 4.4.2. *Let $\delta^* \in \mathcal{D}$. The following are equivalent:*

1. *δ^* is Bayes;*
2. *$\forall \delta \in \mathcal{D}$, $\ell_{\delta^*} \leq \ell_\delta$ μ-a.e.; and*
3. *$\ell_{\delta^*} = \bigwedge \{\ell_\delta : \delta \in \mathcal{D}\}$ μ-a.e.*
4. *For every $\delta \in \mathcal{D}$ and measurable subset $B \subseteq \mathcal{X}$,*

$$\mathbb{E}[\ell_{\delta^*}(\xi); \xi \in B] \leq \mathbb{E}[\ell_\delta(\xi); \xi \in B]. \tag{4.36}$$

Remark 4.4.3. Suppose δ^* is Bayes. If $\mathcal{X}$ is a Borel subset of $\mathbb{R}^n$, then there exists some countable subset $\mathcal{D}' \subseteq \mathcal{D}$ such that $\ell_{\delta^*} = \bigwedge \{\ell_\delta : \delta \in \mathcal{D}\} = \inf_{\delta \in \mathcal{D}'} \ell_\delta$ μ-a.e. [HM02, Lemma 2.6].

We have shown that, if a Bayes procedure δ^* exists, its posterior risk function ℓ_{δ^*} is the lattice infimum (w.r.t. μ) among all posterior risk functions. We now relate this to the more elementary notion of a Bayes procedure being one that minimizes the posterior risk pointwise.

For every $\alpha \in \mathcal{A}$, let $\ell'_\alpha(\cdot) = \ell_{\check{\alpha}}(\cdot)$. That is, $\ell'_\alpha(\cdot) : \mathcal{X} \to \mathbb{R}_+$ is a measurable function satisfying $\ell'_\alpha(\xi) = \mathbb{E}^\xi \bar{L}(\alpha)$.

Definition 4.4.4. We say that *coherent version exists for* $\delta \in \mathcal{D}$ if $\ell'_{\delta(\xi)}(\xi) = \ell_\delta(\xi)$ a.s. We say that *coherent versions exist* if coherent version exists for all $\delta \in \mathcal{D}$.

Next, we discuss sufficient conditions for a decision procedure to be Bayes if coherent versions exist. The proof is given in Section 4.4.2.

Lemma 4.4.5. *Assume coherent versions exist. Let $\delta_0 \in \mathcal{D}$ and assume there exists some μ-measure one set on which $\ell_{\delta_0} = \inf\{\ell'_\alpha(\cdot) : \alpha \in \mathcal{A}\}$. Then δ_0 is Bayes.*

Particularly, the following result shows that if $\mathcal{A}$ is finite, Bayes procedure always exists. The proof is given in Section 4.4.2.

Lemma 4.4.6. *Assume $\mathcal{A}$ is finite. Then there exists $\delta_0 \in \mathcal{D}$ such that $\ell_{\delta_0} = \inf\{\ell'_\alpha(\cdot) : \alpha \in \mathcal{A}\}$. In particular, δ_0 is Bayes.*

For a function f, write $f[A]$ for the image of a set A under f. The following result shows that, a coherent version exists for any δ such that $\delta[\mathcal{X}]$ is countable. The proof is given in Section 4.4.2.

Lemma 4.4.7. *Let $\delta \in \mathcal{D}$ and assume $\delta[\mathcal{X}] \subseteq \mathcal{A}'$ where $\mathcal{A}' \subseteq \mathcal{A}$ is countable. Then a coherent version exists for δ.*

The following is a direct consequence of Lemma 4.4.7.

Corollary 4.4.8. *If $\mathcal{A}$ is countable, coherent versions exist.*

Fix a separable topology on $\mathcal{A}$, i.e., with a countable dense subset. A process $\bar{L}$ indexed by $\mathcal{A}$ is *continuous* if, almost surely, the map $\alpha \mapsto \bar{L}(\alpha)$ is a continuous function from $\mathcal{A}$ to $\mathbb{R}$.

Lemma 4.4.9. *Assume $\mathcal{A}$ is separable and $\bar{L}$ is a continuous process. If $L(\alpha) \leq Y$ a.s. for all $\alpha \in \mathcal{A}$, where Y is integrable, and there exists a sequence of jointly measurable functions $\alpha_n : \Omega \times \mathcal{A} \to D$, where $D \subseteq \mathcal{A}$ be a countable dense subset, such that*

$$\mathbb{P}\left(\forall \alpha, \lim_{n\to\infty} \bar{L}(\alpha_n(\alpha)) = \bar{L}(\alpha)\right) = 1, \tag{4.37}$$

then coherent versions exist.

The proof of Lemma 4.4.9 is given in Section 4.4.2. The following corollary is a direct consequence of Lemma 4.4.9.

Corollary 4.4.10. *Assume $\mathcal{A}$ is separable and $\bar{L}$ is a uniformly continuous process. If $L(\alpha) \leq Y$ a.s. for all $\alpha \in \mathcal{A}$, where Y is integrable, then coherent versions exist.*

4.4.2 Proofs

Proof of Lemma 4.4.1

We have $\bar{L}(\check{\alpha})(\xi) = \bar{L}(\alpha)$, where the r.h.s. is, by definition, a random variable. Similarly, $\bar{L}([\delta_1 : B_1, \delta_2 : B_2, \dots](\xi)) = \sum_{n=1}^{\infty} 1_{\{\xi \in B_n\}} \bar{L}(\delta_n(\xi))$. The r.h.s. is also random variable, being a countable sum of finite products of random variables.

Proof of Lemma 4.4.2

That Item 2 implies Item 1 is immediate from the definition of ℓ_δ and the chain rule.

To see that Item 1 implies Item 2, assume δ^* is Bayes and there exists $\delta \in \mathcal{D}$ such that Eq. (4.38) fails. Then, for some measurable set $B \subseteq \mathcal{X}$, with $\mu(B) > 0$, we have

$$\mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta^*(\xi)); \xi \in B] > \mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta(\xi)); \xi \in B]. \tag{4.38}$$

Let $\bar{B} = \mathcal{X} \setminus B$ and define $\delta' = [\delta : B; \delta^* : \bar{B}]$, which is an element of $\mathcal{D}$ by Lemma 4.4.1. Then

$$\begin{aligned}
\mathbb{E}\bar{L}(\delta'(\xi)) = \mathbb{E}\mathbb{E}^\xi \bar{L}(\delta'(\xi)) &= \mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta(\xi)); \xi \in B] + \mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta^*(\xi)); \xi \in \bar{B}] && (4.39)\\
&< \mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta^*(\xi)); \xi \in B] + \mathbb{E}[\mathbb{E}^\xi \bar{L}(\delta^*(\xi)); \xi \in \bar{B}] && (4.40)\\
&= \mathbb{E}\mathbb{E}^\xi \bar{L}(\delta^*(\xi)) = \mathbb{E}\bar{L}(\delta^*(\xi)), && (4.41)
\end{aligned}$$

a contradiction.

Finally, clearly Item 3 implies Item 2. To see the other direction, note that Item 2 implies that ℓ_{δ^*} is a lower bound on the set $\{\ell_\delta : \delta \in \mathcal{D}\}$, with respect to the partial order $f \succeq g$ iff $f \geq g$ μ-a.e., thus Eq. (4.34) holds. Because ℓ_{δ^*} is also an element of this set, Eq. (4.35) holds, hence ℓ_{δ^*} is the infimum.

For Item 4, assume δ^* is not Bayes, which implies there exists $\delta \in \mathcal{D}$ whose Bayes risk is less. Take $B = \{x \in \mathcal{X} : \ell_{\delta^*}(x) \neq \ell_\delta(x)\}$. This set is measurable. Therefore, Eq. (4.36) and the chain rule implies δ^* has risk no greater than that of δ, a contradiction.

Proof of Lemma 4.4.5

Let $\delta \in \mathcal{D}$. Then, on some μ-measure-one set,

$$\ell_{\delta_0}(\xi) = \inf_{\alpha \in \mathcal{A}} \ell'_\alpha(\xi) \leq \ell'_{\delta(\xi)}(\xi). \tag{4.42}$$

By coherence, $\ell'_{\delta(\xi)}(\xi) = \ell_\delta(\xi)$ a.s. Hence Item 3 of Lemma 4.4.2 holds of δ_0, and so δ_0 is Bayes.

Proof of Lemma 4.4.6

Write $\mathcal{A} = \{\alpha_1, \dots, \alpha_k\}$. For $i \in [k]$, let $B_i \subseteq \mathcal{X}$ be the (measurable) set on which $\alpha_i = \arg_\alpha \inf\{\ell'_\alpha(\cdot) : \alpha \in \mathcal{A}\}$. Then $\{B_1, \dots, B_k\}$ is a partition of $\mathcal{X}$. Let $\delta_0 = [\check{\alpha}_1 : B_1, \dots, \check{\alpha}_k : B_k]$. Then $\delta_0 \in \mathcal{D}$ by Lemma 4.4.1. Clearly, $\ell_{\delta_0} = \inf\{\ell'_\alpha(\cdot) : \alpha \in \mathcal{A}\}$, and so by Lemma 4.4.5, δ_0 is Bayes.

Proof of Lemma 4.4.7

Let $\mathcal{A}' = \{\alpha_1, \alpha_2, \ldots\}$. We can write

$$\bar{L}(\delta(\xi)) = \sum_{\alpha \in \mathcal{A}'} 1_{\{\delta(\xi)=\alpha\}} \bar{L}(\alpha). \tag{4.43}$$

Let $\mathcal{A}'_n = \{\alpha_1, \ldots, \alpha_n\}$ and $\bar{L}(\delta(\xi))_n = \sum_{\alpha \in \mathcal{A}'_n} 1_{\{\delta(\xi)=\alpha\}} \bar{L}(\alpha)$. Since L is nonnegative, we have $\bar{L}(\delta(\xi))_n \leq \bar{L}(\delta(\xi))$ and $\bar{L}(\delta(\xi))_n \to \bar{L}(\delta(\xi))$ a.s. By conditional monotone convergence theorem, we have, almost surely, $\lim_{n\to\infty} \mathbb{E}^\xi \bar{L}(\delta(\xi))_n = \mathbb{E}^\xi \bar{L}(\delta(\xi))$. It follows that, almost surely,

$$\begin{aligned}
\mathbb{E}^\xi \bar{L}(\delta(\xi)) &= \lim_{n\to\infty} \mathbb{E}^\xi \sum_{\alpha \in \mathcal{A}'_n} 1_{\{\delta(\xi)=\alpha\}} \bar{L}(\alpha) && (4.44)\\
&= \lim_{n\to\infty} \sum_{\alpha \in \mathcal{A}'_n} 1_{\{\delta(\xi)=\alpha\}} \mathbb{E}^\xi \bar{L}(\alpha) && (4.45)\\
&= \sum_{\alpha \in \mathcal{A}'} 1_{\{\delta(\xi)=\alpha\}} \mathbb{E}^\xi \bar{L}(\alpha) && (4.46)\\
&= \sum_{\alpha \in \mathcal{A}'} 1_{\{\delta(\xi)=\alpha\}} \ell'_\alpha(\xi) && (4.47)\\
&= \ell'_{\delta(\xi)}(\xi). && (4.48)
\end{aligned}$$

Proof of Lemma 4.4.9

From Eq. (4.37), we have $\bar{L}(\delta(\xi)) = \lim_{n\to\infty} \bar{L}(\alpha_n(\delta(\xi)))$. Since $\bar{L}(\alpha_n(\delta(\xi))) \leq Y$ a.s., by conditional dominated convergence theorem, we have, a.s.,

$$\ell_\delta(\xi) = \mathbb{E}^\xi \bar{L}(\delta(\xi)) = \lim_{n\to\infty} \mathbb{E}^\xi \bar{L}(\alpha_n(\delta(\xi))). \tag{4.49}$$

By the countability of D, a.s.,

$$\bar{L}(\alpha_n(\delta(\xi))) = \sum_{\alpha \in D} 1_{\{\alpha_n(\delta(\xi))=\alpha\}} \bar{L}(\alpha). \tag{4.50}$$

By conditional monotone convergence theorem, almost surely,

$$\begin{aligned}
\mathbb{E}^\xi \bar{L}(\alpha_n(\delta(\xi))) &= \sum_{\alpha \in D} 1_{\{\alpha_n(\delta(\xi))=\alpha\}} \mathbb{E}^\xi \bar{L}(\alpha) && (4.51)\\
&= \sum_{\alpha \in D} 1_{\{\alpha_n(\delta(\xi))=\alpha\}} \ell'_\alpha(\xi) && (4.52)\\
&= \ell'_{\alpha_n(\delta(\xi))}(\xi). && (4.53)
\end{aligned}$$

Furthermore, by Eq. (4.37) and conditional dominated convergence theorem, we have

$$\mathbb{P}\left(\forall \alpha,\ \mathbb{E}^\xi \bar{L}(\alpha) = \lim_{n\to\infty} \mathbb{E}^\xi \bar{L}(\alpha_n(\alpha))\right) = 1. \tag{4.54}$$

Note that $\ell'_\alpha(\xi) = \mathbb{E}^\xi \bar{L}(\alpha)$. Therefore, we have $\lim_{n\to\infty} \ell'_{\alpha_n(\delta(\xi))}(\xi) = \ell'_{\delta(\xi)}(\xi)$ a.s., completing the proof.

Proof of Proposition 4.3.5

From Corollary 4.3.3, we know that, under πQ and log loss, the unique Bayes optimal action is $\pi Q_{\mathcal{Y}|\mathcal{X}}(x)$. Therefore, the meta-Bayesian risk is

$$R(P,\pi) = \int -\log\left(\frac{\mathrm{d}(\pi Q_{\mathcal{Y}|\mathcal{X}}(x))}{\mathrm{d}\nu}(y)\right) P(\mathrm{d}(x,y)) \tag{4.55}$$

$$= \int \left[\int -\log\left(\frac{\mathrm{d}(\pi Q_{\mathcal{Y}|\mathcal{X}}(x))}{\mathrm{d}\nu}(y)\right) P_{\mathcal{Y}|\mathcal{X}}(\mathrm{d}y|x)\right] P^1(\mathrm{d}x). \tag{4.56}$$

By assumption, $P_{\mathcal{Y}|\mathcal{X}}(x) \ll \nu$ for P^1-almost all x. Write $p_2(\cdot|x)$ for the conditional density. By the assumption of finite differential entropy, we have

$$\left|\int \log\left(p_2(y|x)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right| < \infty. \tag{4.57}$$

Then $R(P,\pi)$ can be written as

$$R(P,\pi) = \int \left[\int -\log\left(\frac{\mathrm{d}(\pi Q_{\mathcal{Y}|\mathcal{X}}(x))}{\mathrm{d}\nu}(y)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right] P^1(\mathrm{d}x) \tag{4.58}$$

$$= \int \left[\int -\log\left(\frac{\mathrm{d}(\pi Q_{\mathcal{Y}|\mathcal{X}}(x))}{\mathrm{d}\nu}(y)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right. \tag{4.59}$$

$$\left. - \int \log\left(p_2(y|x)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right. \tag{4.60}$$

$$\left. + \int \log\left(p_2(y|x)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right] P^1(\mathrm{d}x) \tag{4.61}$$

$$= \int \left[\int \log\left(\frac{p_2(y|x)}{\frac{\mathrm{d}(\pi Q_{\mathcal{Y}|\mathcal{X}}(x))}{\mathrm{d}\nu}(y)}\right) p_2(y|x)\,\nu(\mathrm{d}y)\right. \tag{4.62}$$

$$\left. - \int \log\left(p_2(y|x)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right] P^1(\mathrm{d}x) \tag{4.63}$$

$$= \int \left[\mathrm{KL}\left(P_{\mathcal{Y}|\mathcal{X}}(x)||\pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) - \int \log\left(p_2(y|x)\right) p_2(y|x)\,\nu(\mathrm{d}y)\right] P^1(\mathrm{d}x). \tag{4.64}$$

As the second term does not depend on π, if $\pi \in \Pi$, π also achieves the minimum of

$$\min_{\pi\in\mathcal{F}} \int \mathrm{KL}\left(P_{\mathcal{Y}|\mathcal{X}}(x)||\pi Q_{\mathcal{Y}|\mathcal{X}}(x)\right) P^1(\mathrm{d}x). \tag{4.65}$$

Proof of Proposition 4.3.7

By the definition of $R(P,\pi)$, we have

$$R(P,\pi) = \int L(y, A_\pi(x)) P(\mathrm{d}(x,y)),$$

where $L(y,a) = \|a-y\|_2^2$. Since $A_\pi(x) = m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)]$, we have

$$\begin{aligned} R(P,\pi) &= \int \left[\int \|A_\pi(x) - y\|_2^2 P_{\mathcal{Y}|\mathcal{X}}(\mathrm{d}y|x) \right] P^1(\mathrm{d}x) \\ &= \int y^T y P(\mathrm{d}(x,y)) \\ &\quad + \int \left[A_\pi(x)^T A_\pi(x) - A_\pi(x)^T m[P_{\mathcal{Y}|\mathcal{X}}(x)] - m[P_{\mathcal{Y}|\mathcal{X}}(x)]^T A_\pi(x) \right] P^1(\mathrm{d}x), \end{aligned} \tag{4.66}$$

where $m[P_{\mathcal{Y}|\mathcal{X}}(x)]$ is the first moment of $P_{\mathcal{Y}|\mathcal{X}}(\mathrm{d}y|x)$.

Since $m[P_{\mathcal{Y}|\mathcal{X}}(x)]$ does not depend on π, we may replace the first term by $\int m[P_{\mathcal{Y}|\mathcal{X}}(x)]^T m[P_{\mathcal{Y}|\mathcal{X}}(x)] P^1(\mathrm{d}x)$. Then

$$\arg\min_{\pi\in\mathcal{F}} R(P,\pi) = \arg\min_{\pi\in\mathcal{F}} \int \|m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)] - m[P_{\mathcal{Y}|\mathcal{X}}(x)]\|^2 P^1(\mathrm{d}x).$$

i.e., on average of x, the posterior mean using π, $m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)]$ is close to the "true" posterior mean $m[P_{\mathcal{Y}|\mathcal{X}}(x)]$.

Proof of Proposition 4.3.8

For any $\pi \in \mathcal{F}$ and $\pi' \in \tilde{\Pi}$, we have

$$\begin{aligned} R(P,\pi) &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, A_\pi(x)) P(\mathrm{d}x \times \mathrm{d}y) \\ &\geq \int_{\mathcal{X}\times\mathcal{Y}} L(y, A_P(x)) P(\mathrm{d}x \times \mathrm{d}y) \\ &= \int_{\mathcal{X}\times\mathcal{Y}} L(y, A_{\pi'}(x)) P(\mathrm{d}x \times \mathrm{d}y) \\ &= R(P,\pi') \end{aligned} \tag{4.67}$$

Therefore, $\pi' \in \Pi$.

On the other hand, if there exists $\pi'' \in \Pi \setminus \tilde{\Pi}$, then there exists a measurable subset $B \subseteq \mathcal{X}$ such that $P^1(B) > 0$ and for all $x \in B$, $A_P(x) \neq A_{\pi''}(x)$. Then, we have

$$\begin{aligned} &\int_{B\times\mathcal{Y}} L(y, A_{\pi''}(x)) P(\mathrm{d}x \times \mathrm{d}y) \\ &> \int_{B\times\mathcal{Y}} L(y, A_P(x)) P(\mathrm{d}x \times \mathrm{d}y) \\ &= \int_{B\times\mathcal{Y}} L(y, A_{\pi'}(x)) P(\mathrm{d}x \times \mathrm{d}y). \end{aligned} \tag{4.68}$$

Therefore, we have $R(P,\pi') < R(P,\pi'')$, which contradicts $\pi'' \in \Pi$.

Proof of Theorem 4.3.16

We begin with a simple well-known bound:

Lemma 4.4.11. *If $x > 0$, then*

$$1 - \frac{1}{x} \leq \log x \leq x - 1 \tag{4.69}$$

with equality holds when $x = 1$.

For every measurable set $U \subseteq \Theta$, measurable function $h(\cdot \,|\, \cdot) : \mathcal{X} \times \Theta \to \mathbb{R}_+$, and prior $\pi \in \mathcal{M}_1(\Theta)$, define

$$\ell_U(h, \pi) := \int_U \prod_{i=1}^{k} h(x_i|p)\pi(\mathrm{d}p). \tag{4.70}$$

Then

$$\frac{\mathrm{d}(\tilde{\pi}P^{(k)})}{\mathrm{d}\mu} = \int_\Theta \prod_{i=1}^{k} g^*(x_i \,|\, p)\tilde{\pi}(\mathrm{d}p) = \ell_\Theta(g^*, \tilde{\pi}), \quad \frac{\mathrm{d}(\pi Q^k)}{\mathrm{d}\mu} = \int_\Theta \prod_{i=1}^{k} g(x_i \,|\, p)\pi(\mathrm{d}p) = \ell_\Theta(g, \pi). \tag{4.71}$$

Recall the Kullback–Leibler divergence

$$\mathrm{KL}\left(\tilde{\pi}P^{(k)} || \pi Q^k\right) = \int_{\mathcal{X}^k} \log\left(\frac{\mathrm{d}(\tilde{\pi}P^{(k)})}{\mathrm{d}(\pi Q^k)}\right) \tilde{\pi}P^{(k)}(\mathrm{d}x) \tag{4.72}$$

$$= \int_{\mathcal{X}^k} \log\left(\frac{\ell_\Theta(g^*, \tilde{\pi})}{\ell_\Theta(g, \pi)}\right) \tilde{\pi}P^{(k)}(\mathrm{d}x) \tag{4.73}$$

Therefore, it is equivalent to show

$$\mathrm{KL}\left(\tilde{\pi}P^{(k)} || \tilde{\pi}Q^k\right) - \inf_{\pi \in \mathcal{F}} \mathrm{KL}\left(\tilde{\pi}P^{(k)} || \pi Q^k\right) \tag{4.74}$$

$$= \sup_{\pi \in \mathcal{F}} \int_{\mathcal{X}^k} \log\left(\frac{\ell_\Theta(g, \pi)}{\ell_\Theta(g, \tilde{\pi})}\right) \tilde{\pi}P^{(k)}(\mathrm{d}x) \tag{4.75}$$

$$= \sup_{\pi \in \mathcal{F}} \int_\Theta \int_{\mathcal{X}^k} \log\left(\frac{\ell_\Theta(g, \pi)}{\ell_\Theta(g, \tilde{\pi})}\right) P_\theta^{(k)}(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \to 0. \tag{4.76}$$

The key observation is that, by Lemma 4.3.15, we have

$$\lim_{k\to\infty} \frac{\ell_\Theta(g, \pi)}{\ell_\Theta(g, \tilde{\pi})} = \lim_{k\to\infty} \frac{\ell_U(g, \pi)}{\ell_U(g, \tilde{\pi})} \quad a.s. \quad P_\theta \tag{4.77}$$

for any open set U contains $A_0 = \{\theta\}$.

Using the inequality

$$\ell_U(g, \tilde{\pi}) = \int_{U:\theta\in U} \left[\prod_{i=1}^{k} g(x_i \,|\, p) f_{\tilde{\pi}}(p)\right] \left[\frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right] \mathrm{d}p \tag{4.78}$$

$$\leq \left[\sup_{p\in U} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right] \int_{U:\theta\in U} \prod_{i=1}^{k} g(x_i \,|\, p) f_{\tilde{\pi}}(p)\mathrm{d}p = \left[\sup_{p\in U} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right] \ell_U(g, \tilde{\pi}), \tag{4.79}$$

we can get that for all k and any open set $U \subseteq \Theta$, by assumption,

$$\frac{\ell_U(g, \pi)}{\ell_U(g, \tilde{\pi})} \leq \sup_{p\in U} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)} < \infty \quad \text{a.s.} \quad P_\theta^{(k)}. \tag{4.80}$$

Since by assumption, $\frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}$ has a continuous extension to $\bar{\Theta}$, then for any k we can get

$$\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})} \leq \sup_{p\in\bar{\Theta}} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)} < \infty \quad \text{a.s.} \quad P_\theta^{(k)} \tag{4.81}$$

which implies that for any k and θ, we have

$$\log\left(\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) < \infty \quad \text{a.s.} \quad P_\theta^{(k)} \tag{4.82}$$

then

$$\int \log\left(\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) P_\theta^{(k)}(\mathrm{d}x) < \infty. \tag{4.83}$$

This allows us to use the dominated convergence theorem. Letting $U = B_\theta(\epsilon)$, then for any $\epsilon > 0$, we have

$$\lim_k \int\int \log\left(\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) P_\theta^{(k)}(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \tag{4.84}$$

$$\underbrace{=}_{\text{by Eq. (4.83)}} \int \lim_k \int \log\left(\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) P_\theta^{(k)}(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \tag{4.85}$$

$$\underbrace{=}_{\text{by Eq. (4.82)}} \int\int \lim_k \log\left(\frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) P_\theta(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \tag{4.86}$$

$$\underbrace{=}_{\text{by Eq. (4.81)}} \int\int \log\left(\lim_k \frac{\ell_\Theta(g,\pi)}{\ell_\Theta(g,\tilde{\pi})}\right) P_\theta(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \tag{4.87}$$

$$\underbrace{=}_{\text{by Eq. (4.77)}} \int\int \log\left(\lim_k \frac{\ell_{B_\theta(\epsilon)}(g,\pi)}{\ell_{B_\theta(\epsilon)}(g,\tilde{\pi})}\right) P_\theta(\mathrm{d}x)\tilde{\pi}(\mathrm{d}\theta) \tag{4.88}$$

$$\underbrace{\leq}_{\text{by Eq. (4.80)}} \int \log\left(\sup_{p\in B_\theta(\epsilon)} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right) f_{\tilde{\pi}}(\theta)\mathrm{d}\theta \tag{4.89}$$

$$\underbrace{\leq}_{\text{by Lemma 4.4.11}} \int \left[\sup_{p\in B_\theta(\epsilon)} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right] f_{\tilde{\pi}}(\theta)\mathrm{d}\theta - 1. \tag{4.90}$$

which holds for all $\epsilon > 0$.

Since the above upper bound is finite for any ϵ and by continuity the limit

$$\lim_{\epsilon\to 0} \sup_{p\in B_\theta(\epsilon)} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)} \downarrow \frac{f_\pi(\theta)}{f_{\tilde{\pi}}(\theta)} \tag{4.91}$$

for any θ. Finally, by assumption

$$\sup_{p\in B_\theta(\epsilon)} \frac{f_\pi(p)}{f_{\tilde{\pi}}(p)} < \infty, \tag{4.92}$$

by dominated convergence theorem, we have

$$\lim_{\epsilon\to 0}\int\left[\sup_{p\in B_\theta(\epsilon)}\frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right]f_{\tilde{\pi}}(\theta)\mathrm{d}\theta=\int\left[\lim_{\epsilon\to 0}\sup_{p\in B_\theta(\epsilon)}\frac{f_\pi(p)}{f_{\tilde{\pi}}(p)}\right]f_{\tilde{\pi}}(\theta)\mathrm{d}\theta \tag{4.93}$$

$$=\int\frac{f_\pi(\theta)}{f_{\tilde{\pi}}(\theta)}f_{\tilde{\pi}}(\theta)\mathrm{d}\theta=1. \tag{4.94}$$

Proof of Proposition 4.3.17

For i.i.d. Bernoulli case, the optimality condition for the general case reduces to

$$P_{\mathcal{Y}|\mathcal{X}}(\{y\}|x)=\pi Q_{\mathcal{Y}|\mathcal{X}}(\{y\}|x),\quad \forall y\in\mathbb{B}^k,\quad \forall x\in\mathbb{B}^n.$$

If we write $\tilde{\pi}$ as the prior belief, and π as the surrogate prior, we get

$$\frac{\int\theta^{n_1}(1-\theta)^{n_2}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n_1'}(1-\theta)^{n_2'}\tilde{\pi}(\mathrm{d}\theta)}=\frac{\int\theta^{n_1}(1-\theta)^{n_2}\pi(\mathrm{d}\theta)}{\int\theta^{n_1'}(1-\theta)^{n_2'}\pi(\mathrm{d}\theta)},$$

for all nonnegative integers n_1,n_2,n_1',n_2' such that

$$n_1'+n_2'=n,\quad n_1+n_2=n+k,\quad n_1\geq n_1',\quad n_2\geq n_2'.$$

Now for fixing (n_1,n_2), the following equality of $c(n_1,n_2)$ holds for all (n_1',n_2') such that $n_1'\leq n_1,n_2'\leq n_2,n_1'+n_2'=n$.

$$c(n_1,n_2):=\frac{\int\theta^{n_1}(1-\theta)^{n_2}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n_1}(1-\theta)^{n_2}\pi(\mathrm{d}\theta)}=\frac{\int\theta^{n_1'}(1-\theta)^{n_2'}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n_1'}(1-\theta)^{n_2'}\pi(\mathrm{d}\theta)}.$$

Taking $(n_1,n_2)=(n+k-1,1)$ yields

$$c(n+k-1,1)=\frac{\int\theta^{n}(1-\theta)^{0}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n}(1-\theta)^{0}\pi(\mathrm{d}\theta)}=\frac{\int\theta^{n-1}(1-\theta)^{1}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n-1}(1-\theta)^{1}\pi(\mathrm{d}\theta)}.$$

Using the notation $m_k[\tilde{\pi}]$ and $m_k[\pi]$, the above equation implies

$$c(n+k-1,1)=\frac{m_n[\tilde{\pi}]}{m_n[\pi]}=\frac{m_{n-1}[\tilde{\pi}]-m_n[\tilde{\pi}]}{m_{n-1}[\pi]-m_n[\pi]}.$$

Therefore, we have

$$\frac{m_n[\tilde{\pi}]}{m_n[\pi]}=\frac{m_{n-1}[\tilde{\pi}]}{m_{n-1}[\pi]}.$$

Similarly, if we continue to write down $c(n+k-2,2),\cdots,c(0,n+k)$, eventually we can get

$$\frac{m_n[\tilde{\pi}]}{m_n[\pi]}=\frac{m_{n-1}[\tilde{\pi}]}{m_{n-1}[\pi]}=\cdots=\frac{m[\tilde{\pi}]}{m[\pi]}=1.$$

Next, we fix (n_1',n_2') and consider

$$c'(n_1',n_2'):=\frac{\int\theta^{n_1'}(1-\theta)^{n_2'}\tilde{\pi}(\mathrm{d}\theta)}{\int\theta^{n_1'}(1-\theta)^{n_2'}\pi(\mathrm{d}\theta)}=c(n_1,n_2),\quad \forall n_1\geq n_1',\quad n_2\geq n_2',\quad n_1+n_2=n+k.$$

Taking $(n_1', n_2') = (n, 0)$, we have

$$c'(n,0) = c(n,k) = c(n+1,k-1) = \cdots = c(n+k,0),$$

which lead to

$$\frac{m_{n+k}[\tilde{\pi}]}{m_{n+k}[\pi]} = \frac{m_{n+k-1}[\tilde{\pi}]}{m_{n+k-1}[\pi]} = \cdots = \frac{m_{n+1}[\tilde{\pi}]}{m_{n+1}[\pi]} = 1.$$

Therefore, we have the result that in i.i.d Bernoulli case, matching the first $n+k$ moments is optimal (if feasible).

Proof of Proposition 4.3.18

Since both $P_{\mathcal{Y}|\mathcal{X}}(x)$ and $\pi Q_{\mathcal{Y}|\mathcal{X}}(x)$ are exchangeable measures on $\mathbb{B}^k$, the posterior means $m[P_{\mathcal{Y}|\mathcal{X}}(x)]$ and $m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)]$ are vectors of length k whose elements are equal. Therefore, it suffices to consider $k = 1$. The property

$$m[P_{\mathcal{Y}|\mathcal{X}}(x)] = m[\pi Q_{\mathcal{Y}|\mathcal{X}}(x)], \quad \forall x \in \mathbb{B}^n, \tag{4.95}$$

is equivalent to

$$P_{\mathcal{Y}|\mathcal{X}}(\{y\}|x) = \pi Q_{\mathcal{Y}|\mathcal{X}}(\{y\}|x), \quad \forall y \in \mathbb{B}^1, \quad \forall x \in \mathbb{B}^n. \tag{4.96}$$

According to Proposition 4.3.17, we can get $m_i[\pi] = m_i[\tilde{\pi}], \quad \forall i = 1, \ldots, n+1$.

Proof of Theorem 4.3.19

We begin with a well-known lemma.

Lemma 4.4.12 (Laplace approximation)**.** *Let S and f be continuous functions on $(0,1)$, assume that the supremum of S is achieved at a unique point $x_0 \in (0,1)$, that S is three-times continuously differentiable in a neighborhood of x_0, and that $S''(x_0) \neq 0$. Let*

$$F(\lambda) = \int_0^1 f(x) \exp(\lambda S(x)) \mathrm{d}x, \quad \textit{for } \lambda \in \mathbb{R}. \tag{4.97}$$

If it is known that $f(x_0) \neq 0$ and $f(x) = f(x_0) + \mathcal{O}(x - x_0)$ for $x \to x_0$, then, as $\lambda \to \infty$, we have

$$F(\lambda) = \sqrt{\frac{2\pi}{-S''(x_0)}} f(x_0) \exp(\lambda S(x_0)) \lambda^{-1/2} [1 + \mathcal{O}(\lambda^{-1/2})]. \tag{4.98}$$

Proof. See [ZC04, Thm. 1, Chp. 19.2.4]. □

Under the assumption of finite differential entropy, it is straightforward to establish that

$$\begin{aligned} &R(\tilde{\pi}P^{(k)}, \tilde{\pi}) - \inf_{\pi \in \mathcal{M}_1^c[0,1]} R(\tilde{\pi}P^{(k)}, \pi) \\ &\quad = \mathrm{KL}\left(\tilde{\pi}P^{(k)} || \tilde{\pi}Q^k\right) - \inf_{\pi \in \mathcal{M}_1^c[0,1]} \mathrm{KL}\left(\tilde{\pi}P^{(k)} || \pi Q^k\right). \end{aligned} \tag{4.99}$$

Thus it suffices to show

$$\sup_{\pi} \int \int \log \left(\frac{\pi Q^k\{x\}}{\tilde{\pi} Q^k\{x\}} \right) P_\theta^{(k)}(\mathrm{d}x) \tilde{\pi}(\mathrm{d}\theta) \to 0, \tag{4.100}$$

where this and all subsequent limits are taking as $k \to \infty$, unless stated otherwise. First note that $P_\theta^{(k)}$ is a discrete probability measure on $\mathbb{B}^k$ for each k. Let $\tilde{\theta}_{x,k} = \frac{1}{k} \sum_{i=1}^k x_i$ be the relative frequency of 1's in x up to entry k. Then, $k\tilde{\theta}_{x,k}$ is the number of 1's in x, and $k(1-\tilde{\theta}_{x,k})$ is the number of 0's in x, up to entry k. Thus it suffices to show

$$\sup_{\pi} \int \left[\sum_{x \in \mathbb{B}^k} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}\{x\} \right] \tilde{\pi}(\mathrm{d}\theta) \to 0. \tag{4.101}$$

Let $\pi \in \mathcal{M}_1^c[0,1]$ and suppose $\epsilon \in (0, 1/2)$. For every $\theta \in (\epsilon, 1-\epsilon)$, define

$$A_{\epsilon,\theta,k} = \{x \in \mathbb{B}^k : |\tilde{\theta}_{x,k} - \theta| \leq \epsilon\} \tag{4.102}$$

Then since we have assumed that, uniformly over $\theta \in (\epsilon, 1-\epsilon)$, $\tilde{\theta}_{x,k} \to \theta$ a.s. under $P_\theta^{(\infty)}$, we have

$$P_\theta^{(k)}(A_{\epsilon,\theta,k}^c) = P_\theta^{(k)}(\{x \in \mathbb{B}^k : |\tilde{\theta}_{x,k} - \theta| > \epsilon\}) \to 0, \tag{4.103}$$

uniformly over $\theta \in (\epsilon, 1-\epsilon)$. Then, by writing $\mathbb{B}^k = A_{\epsilon,\theta,k} \cup A_{\epsilon,\theta,k}^c$, one can write

$$\begin{aligned}
&\sum_{x \in \mathbb{B}^k} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}(x) \\
&= \sum_{x \in A_{\epsilon,\theta,k}} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}(x) \\
&\quad + \sum_{x \in A_{\epsilon,\theta,k}^c} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}(x).
\end{aligned} \tag{4.104}$$

Let f_π and $f_{\tilde{\pi}}$ be probability density functions for π and $\tilde{\pi}$, respectively, which exist because $\pi, \tilde{\pi} \in \mathcal{M}_1^c[0,1]$ by assumption. By a Laplace approximation (Lemma 4.4.12), we have

$$\begin{aligned}
\int p^{k\theta} (1-p)^{k(1-\theta)} \pi(\mathrm{d}p) &= \int [f_\pi(p)] \exp(k[\theta \log p + (1-\theta)\log(1-p)]) \, \mathrm{d}p \\
&= \sqrt{\frac{2\pi\theta(1-\theta)}{k}} [(\theta)^\theta (1-\theta)^{1-\theta}]^k f_\pi(\theta)[1 + \mathcal{O}(k^{-1/2})].
\end{aligned} \tag{4.105}$$

Note that $f_\pi(\theta)$ has a continuous extension to $[0,1]$ and the extension of $f_{\tilde{\pi}}$ to $[0,1]$ is strictly positive, then $\frac{f_\pi(\theta)}{f_{\tilde{\pi}}(\theta)}$ has a continuous extension to $[0,1]$ so

$$\sup_{\theta^* \in \{|\theta^* - \theta| > \epsilon\}} \frac{f_\pi(\theta^*)}{f_{\tilde{\pi}}(\theta^*)} < \infty, \qquad \sup_{\theta^* \in \{|\theta^* - \theta| \leq \epsilon\}} \frac{f_\pi(\theta^*)}{f_{\tilde{\pi}}(\theta^*)} < \infty. \tag{4.106}$$

Furthermore, by examining the use of Taylor expansion in the proof of Lemma 4.4.12, the constant for the term $\mathcal{O}(\lambda^{-1/2})$ in Lemma 4.4.12 only involves x_0 by the term $S''(x_0)$. In our context of Eq. (4.105),

$S''(x_0)$ equals to $\frac{1}{\theta(1-\theta)}$ and it is bounded for $\theta \in (\epsilon, 1-\epsilon)$. Therefore, following the proof of Lemma 4.4.12 one can get that the term $\mathcal{O}(k^{-1/2})$ in Eq. (4.105) is uniform over $\theta \in (\epsilon, 1-\epsilon)$.

Using Eqs. (4.104) and (4.105) and the Taylor expansion

$$\log(1 + \mathcal{O}(k^{-1/2})) = \mathcal{O}(k^{-1/2}), \tag{4.107}$$

for $\theta \in (\epsilon, 1-\epsilon)$, we have

$$\begin{aligned}
&\sum_{x \in \mathbb{B}^k} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}(x) \\
&\leq P_\theta^{(k)}(A_{\epsilon,\theta,k}) \sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \left[\log \left(\frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right) + \mathcal{O}(k^{-1/2}) \right] \\
&\quad + P_\theta^{(k)}(A^c_{\epsilon,\theta,k}) \sup_{\{\theta^* \in [0,1] : |\theta^* - \theta| > \epsilon\}} \left[\log \left(\frac{f_\pi(\theta^*)}{f_{\tilde{\pi}}(\theta^*)} \right) + \mathcal{O}(k^{-1/2}) \right].
\end{aligned} \tag{4.108}$$

By the facts that the term $\mathcal{O}(k^{-1/2})$ and $P_\theta^{(k)}(A^c_{\epsilon,\theta,k}) \to 0$ are uniform over $\theta \in (\epsilon, 1-\epsilon)$, the second term of the right hand side of Eq. (4.108) goes to zero and the first term

$$P_\theta^{(k)}(A_{\epsilon,\theta,k}) \sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \left[\log \left(\frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right) + \mathcal{O}(k^{-1/2}) \right] \to \sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \log \left(\frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right) < \infty, \tag{4.109}$$

both uniformly over $\theta \in (\epsilon, 1-\epsilon)$. Then, as $k \to \infty$, using Lemma 4.4.11, we have

$$\begin{aligned}
&\lim_{k \to \infty} \int \left[\sum_{x \in \mathbb{B}^k} \log \left(\frac{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \pi(\mathrm{d}p)}{\int p^{k\tilde{\theta}_{x,k}} (1-p)^{k(1-\tilde{\theta}_{x,k})} \tilde{\pi}(\mathrm{d}p)} \right) P_\theta^{(k)}(x) \right] \tilde{\pi}(\mathrm{d}\theta) \\
&= \int_\epsilon^{1-\epsilon} \left[\sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \log \left(\frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right) \right] f_{\tilde{\pi}}(\theta) \mathrm{d}\theta + \mathcal{O}(\epsilon).
\end{aligned} \tag{4.110}$$

Note that

$$\begin{aligned}
&\int_\epsilon^{1-\epsilon} \left[\sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \log \left(\frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right) \right] f_{\tilde{\pi}}(\theta) \mathrm{d}\theta \\
&\leq \int_\epsilon^{1-\epsilon} \left[\sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} - 1 \right] f_{\tilde{\pi}}(\theta) \mathrm{d}\theta \\
&= \int_\epsilon^{1-\epsilon} \left[\sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right] f_{\tilde{\pi}}(\theta) \mathrm{d}\theta - 1 + \mathcal{O}(\epsilon) < \infty.
\end{aligned} \tag{4.111}$$

As the inequality holds for any $\epsilon \in (0, 1/2)$, letting $\epsilon \downarrow 0$ yields

$$\lim_{\{\epsilon \downarrow 0\}} \int_\epsilon^{1-\epsilon} \left[\sup_{\{\tilde{\theta} \in [0,1] : |\tilde{\theta} - \theta| \leq \epsilon\}} \frac{f_\pi(\tilde{\theta})}{f_{\tilde{\pi}}(\tilde{\theta})} \right] f_{\tilde{\pi}}(\theta) \mathrm{d}\theta - 1 + \mathcal{O}(\epsilon) = \int_0^1 f_\pi(\theta) \mathrm{d}\theta - 1 = 0. \tag{4.112}$$

Chapter 5

Fast-rate PAC-Bayes generalization bounds via shifted Rademacher processes

The developments of Rademacher complexity and PAC-Bayesian theory have been largely independent. One exception is the PAC-Bayes theorem of Kakade, Sridharan, and Tewari [KST08], which is established via Rademacher complexity theory by viewing Gibbs classifiers as linear operators. The goal of this chapter is to extend this bridge between Rademacher complexity and state-of-the-art PAC-Bayesian theory. We first demonstrate that one can match the fast rate of Catoni's PAC-Bayes bounds [Cat07] using shifted Rademacher processes [Weg03; LM12; ZH18]. We then derive a new fast-rate PAC-Bayes bound in terms of the "flatness" of the empirical risk surface on which the posterior concentrates. Our analysis establishes a new framework for deriving fast-rate PAC-Bayes bounds and yields new insights on PAC-Bayesian theory.

5.1 Background

Let $\mathcal{D}$ be an unknown distribution over a space $\mathcal{Z}$ of labeled examples, and let $\mathcal{H}$ be a hypothesis class. Relative to a binary loss function $\ell : \mathcal{H} \times \mathcal{Z} \to \{0,1\}$, we define the associated loss class $\mathcal{F} := \{\ell(h,\cdot) : h \in \mathcal{H}\}$ of functions from $\mathcal{Z} \to \{0,1\}$, each associated to one or more hypotheses. Let $\mathcal{L}_{\mathcal{D}}(f) := \mathbb{E}_{z\sim\mathcal{D}} f(z)$ denote the expected loss, i.e., risk, of every hypothesis associated to f. Let $S = (z_1, \cdots, z_m) \sim \mathcal{D}^m$ be a sequence of i.i.d. random variables. Let $\hat{\mathcal{L}}_S(f) = \frac{1}{m}\sum_{i=1}^m f(z_i)$ denote the empirical risk of every hypothesis associated to f.

We will be primarily interested in Gibbs classifiers, i.e., distributions P on $\mathcal{F}$ which are interpreted as randomized classifiers that classify each new example according to a hypothesis drawn independently from P. (It is more common to work with distributions over $\mathcal{H}$, but these lead to looser results.) For a Gibbs classifier P and labeled example $z \in \mathcal{Z}$, let $\mathbb{E}_P f(z) = \mathbb{E}_{f\sim P}[f(z)]$ be the expected loss P suffers when labeling z. For Gibbs classifiers, the (expected) risk is defined to be $\mathcal{L}_{\mathcal{D}}(P) := \mathbb{E}_{f\sim P}\mathcal{L}_{\mathcal{D}}(f) = \mathbb{E}_{z\sim\mathcal{D}}\mathbb{E}_P f(z)$. The (expected) empirical risk is $\hat{\mathcal{L}}_S(P) := \mathbb{E}_{f\sim P}\hat{\mathcal{L}}_S(f) = \frac{1}{m}\sum_{i=1}^m \mathbb{E}_P f(z_i)$.

5.1.1 PAC-Bayes

The PAC-Bayes framework [McA99] provides data-dependent generalization guarantees for Gibbs classifiers. Each bound is specified in terms of a Gibbs classifier P called the *prior*, as it must be independent of the training sample. The bound then holds for all *posterior* distributions, i.e., Gibbs classifiers that may be defined in terms of the training sample.

Theorem 5.1.1 (PAC-Bayes [McA99])**.** *For any prior distribution P over $\mathcal{F}$, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over draws of training data $S \sim \mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \leq \hat{\mathcal{L}}_S(Q) + \sqrt{\frac{\mathrm{KL}\left(Q||P\right) + \log\frac{m}{\delta}}{2(m-1)}}. \tag{5.1}$$

Note in Theorem 5.1.1, the generalization bound scales as $\mathcal{O}(m^{-\frac{1}{2}})$. Catoni [Cat07] presents a *fast rate* PAC-Bayesian bound, in which the generalization bound scales as $\mathcal{O}(m^{-1})$.

Theorem 5.1.2 (Fast-Rate PAC-Bayes [Cat07, Thm 1.2.6])**.** *For any prior distribution P over $\mathcal{F}$, for any $\delta \in (0,1)$ and $C > 0$, with probability at least $1-\delta$ over draws of training data $S \sim \mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q) \leq \frac{1}{1-e^{-C}}\left[C\hat{\mathcal{L}}_S(Q) + \frac{\mathrm{KL}\left(Q||P\right) + \log\frac{1}{\delta}}{m}\right]. \tag{5.2}$$

Because the constant $C/(1-e^{-C}) > 1$ holds for any $C > 0$, the generalization bound in Theorem 5.1.2 will always be bounded below by the empirical risk. Usually for a well-trained distribution Q over training set, the empirical risk $\hat{\mathcal{L}}_S(Q)$ is small, therefore the generalization bound is dominated by the KL term. Compared to the standard PAC-Bayes bound in Theorem 5.1.1, where the KL term decreases at a rate $\mathcal{O}(m^{-\frac{1}{2}})$, the KL term of Catoni's bound decreases at a rate $\mathcal{O}(m^{-1})$. For this reason, we say that Catoni's bound achieves a *fast* rate of convergence. Note that fast-rate bounds can lead to much tighter bounds. Of course, $C/(1-e^{-C}) \to 1$ as $C \to 0$, but, in that limit, the constants ignored in the asymptotic rate $\mathcal{O}(m^{-1})$ degrade. (See [LLST13] for more discussion.)

5.1.2 Rademacher viewpoint

Fix a prior Gibbs classifier P on $\mathcal{F}$. Then, for measurable functions g, h, consider the inner product $\langle g, h\rangle = \int g(f)h(f)P(\mathrm{d}f)$. The key observation of Kakade, Sridharan, and Tewari is that one can view $\mathcal{L}_{\mathcal{D}}(Q)$ (resp., $\hat{\mathcal{L}}_S(Q)$) as the inner product $\langle \mathrm{d}Q/\mathrm{d}P, \mathcal{L}_{\mathcal{D}}(\cdot)\rangle$ (resp., $\langle \mathrm{d}Q/\mathrm{d}P, \hat{\mathcal{L}}_S(\cdot)\rangle$) between the posterior Q, represented by its Radon–Nikodym derivative with P, and the risk (resp., empirical risk), viewed as measurable function on $\mathcal{F}$. Thus, Gibbs classifiers can be viewed as linear predictors. Using their distribution-independent bounds on the Rademacher complexity of certain classes of linear predictors, Kakade, Sridharan, and Tewari [KST08] derive a PAC-Bayes bound similar to Theorem 5.1.1. We refer to this as the "Rademacher viewpoint" on PAC-Bayes.

We now summarize their argument in more detail. Let $\mathcal{Q}(\kappa) := \{Q : \mathrm{KL}\left(Q||P\right) \leq \kappa\}$. One can follow the classical steps for controlling the generalization error uniformly over $\mathcal{Q}(\kappa)$ using Rademacher complexity. Their first step is to connect $\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q)\right]$ to $\mathbb{E}_S \sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q)\right]$ by the bounded difference inequality (McDiarmid's inequality). In particular, with probability at least

$1-\delta$,

$$\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q)-\hat{\mathcal{L}}_S(Q)\right]\leq\mathbb{E}_S\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q)-\hat{\mathcal{L}}_S(Q)\right]+\sqrt{\frac{\log(1/\delta)}{m}}. \tag{5.3}$$

Then they apply a symmetrization argument to obtain an upper bound in terms of Rademacher complexity [BM02]. In particular, recalling that $S=(z_1,\cdots,z_m)$ is our training data,

$$\mathbb{E}_S\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q)-\hat{\mathcal{L}}_S(Q)\right]\leq 2\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{Q\in\mathcal{Q}(\kappa)}\left[\frac{1}{m}\sum_{i=1}^m\epsilon_i\mathbb{E}_Q f(z_i)\right], \tag{5.4}$$

where $\{\epsilon_i\}$ are i.i.d. Rademacher random variables, i.e., $\mathbb{P}(\epsilon_i=+1)=\mathbb{P}(\epsilon_i=-1)=1/2$. Their last step is to bound the Rademacher complexity $\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\sup_{Q\in\mathcal{Q}(\kappa)}\left[\frac{1}{m}\sum_{i=1}^m\epsilon_i\mathbb{E}_Q f(z_i)\right]$, which can be seen as the Rademacher complexity of a linear class with a (strongly) convex constraint [KST08]. According to [KST08], the Rademacher complexity in Eq. (5.4) is of order $\sqrt{\kappa/m}$, which eventually leads to a term of order $\sqrt{\mathrm{KL}\left(Q||P\right)/m}$ after applying a union bound argument on κ.

In the end, using the above arguments and their sharp bounds on the Rademacher and Gaussian complexities of (constrained) linear classes [KST08, Thm. 1], Kakade, Sridharan, and Tewari obtain the following PAC-Bayes bound [KST08, Cor. 8]: for every prior P over $\mathcal{F}$, with probability at least $1-\delta$ over draws of training data $S\sim\mathcal{D}^m$, for all distribution Q over $\mathcal{F}$,

$$\mathcal{L}_{\mathcal{D}}(Q)\leq\hat{\mathcal{L}}_S(Q)+4.5\sqrt{\frac{\max\{\mathrm{KL}\left(Q||P\right),2\}}{m}}+\sqrt{\frac{\log(1/\delta)}{m}}. \tag{5.5}$$

Note that this PAC-Bayes bound has a slow rate of $\sqrt{1/m}$, but it slightly improves the rate in the term $\sqrt{\log(m/\delta)/m}$ of McAllester's bound [McA99] to $\sqrt{\log(1/\delta)/m}$.

Since McAllester's bound is far from the state-of-art in PAC-Bayesian theory, this raises the question whether one can extend the "Rademacher viewpoint" of PAC-Bayes to derive more advanced bounds, such as one matching the fast rate of Catoni's bound.

5.2 Extending the Rademacher viewpoint

There are at least two difficulties in the "Rademacher viewpoint" that prevent fast rates. First, if we connect the generalization error to Rademacher complexity using the bounded difference inequality, a slow rate term $\sqrt{\log(1/\delta)/m}$ will appear. Second, as is shown by Kakade, Sridharan, and Tewari [KST08], the standard Rademacher complexity of (constraint) linear classes leads to an upper bound with a slow rate of order $\mathcal{O}(\sqrt{\mathrm{KL}\left(Q||P\right)/m})$. Therefore, in order to derive fast rate PAC-Bayes bounds, we need to extend the "Rademacher viewpoint".

In order to obtain fast rates, we work with so-called shifted Rademacher processes, i.e., processes of the form $\{\frac{1}{m}\sum_{i=1}^m\epsilon_i' f(z_i)\}_{f\in\mathcal{F}}$ where the variables $\{\epsilon_i'\}$ are independent from S, i.i.d., and take two values with equal probability. (These shifted Rademacher variables, $\{\epsilon_i'\}$, are not necessarily zero mean. When they take values in $\{\pm1\}$, we obtain a standard Rademacher process.) Shifted Rademacher processes are examples of shifted empirical processes [Weg03; LM12; ZH18].

Recall that Rademacher complexity is the *expected value* of the supremum of Rademacher processes over a class [BM02]. In order to get a fast rate, we connect the *tail probabilities* of the supremum of the

generalization error to the *tail probabilities* of shifted Rademacher processes via a symmetrization-in-deviation argument instead of the symmetrization-in-expectation argument. The key is that we can avoid using the bounded difference inequality by bounding the deviation. This removes the slow rate term of $\sqrt{\log(1/\delta)/m}$. It remains to bound the deviation of shifted Rademacher processes to get a fast rate bound of order $\mathcal{O}(\mathrm{KL}\left(Q||P\right)/m)$.

In the following, we demonstrate how the extended "Rademacher viewpoint" via shifted Rademacher processes can be applied to derive a fast rate PAC-Bayes bound that matches the fast rate of Catoni's bound. Note that, since $C/(1-e^{-C}) > 1$ for fixed $C > 0$ in Catoni's bound in Eq. (5.2), we can write $C/(1-e^{-C}) = 1 + c$ for some constant $c > 0$. Furthermore, note that our goal in this section is not to derive new PAC-Bayes bounds. Therefore, we do not make attempts to optimize the constants.

Proposition 5.2.1 (Matching Catoni's fast rate via shifted Rademacher processes)**.** *For any given $c > 0$ and prior P over $\mathcal{F}$, there exists constants C_1, C_2, and C_3 such that, with probability at least $1-\delta$, for all distributions Q over $\mathcal{F}$*

$$\mathcal{L}_{\mathcal{D}}(Q) \leq (1+c)\hat{\mathcal{L}}_S(Q) + C_1\frac{\mathrm{KL}\left(Q||P\right)}{m} + C_2\frac{\log\frac{1}{\delta}}{m} + C_3\frac{1}{m}. \tag{5.6}$$

Outline of the proof. We wish to emphasize two key differences from traditional machinery for deriving Rademacher-complexity-based generalization bounds. The complete proof is given in Section 5.5.1.

Fix P and let $\mathcal{Q}(\kappa) := \{Q : \mathrm{KL}\left(Q||P\right) \leq \kappa\}$ be defined as in Section 5.1.2. Rather than control $\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q)\right]$ in terms of its *expectation* via the bounded difference inequality and Rademacher complexity, we bound the *tail/deviation* of $\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q)\right]$, thus avoiding the use of the bounded differences inequality altogether. In particular, we can obtain fast rates by bounding the tail in terms of tail of supremum of shifted Rademacher processes [Weg03; LM12; ZH18].

Define $\mathcal{G}_\kappa := \{\mathbb{E}_Q f(\cdot) : Q \in \mathcal{Q}(\kappa)\}$ and, by an abuse of notation, let $\mathcal{L}_{\mathcal{D}}(g)$ denote $\mathbb{E}_{z\sim\mathcal{D}}[g(z)]$. Then we can write $\sup_{Q\in\mathcal{Q}(\kappa)}\left[\mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q)\right]$ as $\sup_{g\in\mathcal{G}_\kappa}\left[\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g)\right]$. We start from bounding the tail probability $\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa}\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) \geq t\right)$. For fixed constants $c > c_2 > 0$, let $c' = \frac{c-c_2}{1+c_2}$ and $t' = \frac{t}{2(1+c_2)}$. Then, by [ZH18, Cor. 1], we have

$$\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa}\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) \geq t\right) \leq 4\mathbb{P}_{S,\epsilon}\left(\sup_{g\in\mathcal{G}_\kappa}\left[\frac{1+\frac{c'}{2}}{m}\sum_{i=1}^m\left(\epsilon_i - \frac{c'}{2+c'}\right)g(z_i)\right] \geq \frac{t'}{2}\right). \tag{5.7}$$

Letting $\epsilon_i' := \epsilon_i - \frac{c'}{2+c'}$, one can see that $\{\epsilon_i'\}$ are i.i.d. "shifted" Rademacher random variables with mean $-\frac{c'}{2+c'}$. For any $g \in \mathcal{G}_\kappa$, there exists $Q \in \mathcal{Q}(\kappa)$ such that

$$\frac{1}{m}\sum_{i=1}^m \epsilon_i' g(z_i) = \frac{1}{m}\sum_{i=1}^m \epsilon_i' \mathbb{E}_Q f(z_i) = \mathbb{E}_Q\left[\frac{1}{m}\sum_{i=1}^m \epsilon_i' f(z_i)\right], \tag{5.8}$$

which can be viewed as a linear function of Q. Further, it can be verified that the set $\mathcal{Q}(\kappa)$ is (strongly) convex. Therefore, $\sup_{Q\in\mathcal{Q}(\kappa)}\frac{1}{m}\sum_{i=1}^m \epsilon_i'\mathbb{E}_Q f(z_i)$ is a convex optimization problem. By duality [BV04, Chp. 5], and, in this particular case, the Legendre transform of Kullback–Leibler divergence (see, e.g.,

[Gue19]), we have

$$\sup_{g\in\mathcal{G}_\kappa}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i' g(z_i)=\sup_{Q\in\mathcal{Q}(\kappa)}\frac{1}{m}\sum_{i=1}^{m}\epsilon_i'\mathbb{E}_Q f(z_i)=\inf_{\lambda>0}\left\{\frac{\kappa}{\lambda}+\frac{1}{\lambda}\log\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right]\right\}. \tag{5.9}$$

Combining the shifted symmetrization in deviation in Eq. (5.7) and the dual problem in Eq. (5.9), Markov's inequality yields, for every $\lambda>0$,

$$\mathbb{P}_S\left(\sup_{Q\in\mathcal{Q}(\kappa)}\mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f)-(1+c)\hat{\mathcal{L}}_S(f)]\geq t\right)\leq 4e^{\kappa-\frac{\lambda t'}{2+c'}}\mathbb{E}_S\mathbb{E}_\epsilon\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right]. \tag{5.10}$$

We then exploit the shifted property of ϵ_i' to bound the expectation term on the right-hand side and obtain fast rates. In particular, we show that, so long as $k\geq\frac{\log\cosh(\lambda/m)}{\lambda/m}$,

$$\mathbb{E}_P\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}(\epsilon_i-k)f(z_i)\right)\right]\leq 1. \tag{5.11}$$

In our case, $k=\frac{c'}{2+c'}$, which leads to constraints relating λ, c, and c_2. In particular, when $c=0$, the required condition for the above result, $k\geq\frac{\log\cosh(\lambda/m)}{\lambda/m}$, does not hold. Therefore, this approach obtains fast rates only if $c>0$, i.e., if we shift. Combing Eqs. (5.10) and (5.11), there exists a constant C', depending only on c, c_2 and δ, such that, with probability at least $1-\delta$,

$$\sup_{Q\in\mathcal{Q}(\kappa)}\mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f)-(1+c)\hat{\mathcal{L}}_S(f)]\leq\frac{C'}{m}(\kappa+\log(4/\delta)). \tag{5.12}$$

Finally, we may apply the same union-bound argument as in the proof of [KST08, Cor. 7] in order to cover all possible values of κ. This completes the proof. □

5.3 New fast-rate PAC-Bayes bound based on "flatness"

The extended " Rademacher viewpoint" of PAC-Bayes provides a new approach for deriving fast-rate PAC-Bayes bounds. In this section, we demonstrate the use of shifted Rademacher processes to derive a new fast-rate PAC-Bayes bound using a notion of "flatness". This notion is inspired by the proposal by Dziugaite and Roy [DR17] to formalize the empirical connection between "flat minima" and generalization using PAC-Bayes bounds, and, in particular, posterior distributions which concentrate in these "flat minima".

Definition 5.3.1 (Notion of "flatness")**.** For given $h\in[0,1]$, the "h-flatness" of Q (w.r.t. the training data S) is

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2. \tag{5.13}$$

One way to understand this new notion is to observe that, under zero–one loss, h-flatness can be

written as the difference between the empirical risk and the quadratic empirical risk:

$$\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2=\hat{\mathcal{L}}_S(Q)-\frac{1-h^2}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2. \tag{5.14}$$

Note that, for $[0,1]$-valued (bounded) loss, equality is replaced by an inequality: the r.h.s. is an upper bound of the l.h.s.

Remark 5.3.2. To see that optimizing h-flatness prefers "flat minima", consider the following simplified case: Call a posterior Q "completely flat" if $f=g$ on S a.s., when $f,g\sim Q$. It can be verified that, if the posterior is "completely flat", then under the zero–one loss, the "h-flatness" is $h^2\hat{\mathcal{L}}_S(Q)$. That is, given a "completely flat" posterior, the "h-flatness" goes to zero as $h\to 0$. For $h>0$, the "h-flatness" is zero when Q is "completely flat" and $\hat{\mathcal{L}}_S(Q)=0$.

The following PAC-Bayes theorem establishes favorable bounds for h-flat posteriors:

Theorem 5.3.3 (Fast-rate PAC-Bayes using "flatness")**.** *For any given $c>0$ and $h\in(0,1)$, with probability at least $1-\delta$ over random draws of training set $S\sim\mathcal{D}^m$, for all distributions Q over $\mathcal{F}$,*

$$\mathcal{L}_{\mathcal{D}}(Q)\leq\hat{\mathcal{L}}_S(Q)+\frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2+\frac{4}{Cm}\left[3\mathrm{KL}\left(Q||P\right)+\log\frac{1}{\delta}+5\right], \tag{5.15}$$

where $C=\frac{2h^4c}{1+16h^2c}$.

This bound can be tighter than Catoni's bound under certain conditions. We delay the comparison with Catoni's bound to Section 5.3.1. We now give an outline of the proof of Theorem 5.3.3, highlighting the technical differences from the proof of Proposition 5.2.1. The complete proof is given in Section 5.5.2.

Outline of the proof of Theorem 5.3.3. By Eq. (5.14), we can write

$$\begin{aligned}&\mathbb{E}_Q\mathcal{L}_{\mathcal{D}}(f)-\hat{\mathcal{L}}_S(Q)-\frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2\\&=\mathcal{L}_{\mathcal{D}}(Q)-(1+c)\hat{\mathcal{L}}_S(Q)+\frac{c(1-h^2)}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2.\end{aligned} \tag{5.16}$$

There are at least two new challenges compared with the proof of Proposition 5.2.1. First, the shifted symmetrization in Eq. (5.7) cannot be applied because of the existence of the quadratic term $\frac{c(1-h^2)}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2$. This means we need to derive a new shifted symmetrization involving the quadratic term. Second, the quadratic term $\frac{c(1-h^2)}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2$ cannot be seen as a linear function of Q. Therefore, some technical arguments are required in order to apply the Legendre transform of Kullback–Leibler divergence.

First, we derive a new shifted symmetrization which involves quadratic terms. The proof is inspired by an argument due to Zhivotovskiy and Hanneke [ZH18]. The result extends [ZH18, Cor. 1], which is recovered as a special case when $h=1$. For $\kappa>0$, recall that we have defined $\mathcal{Q}(\kappa)=\{Q:\mathrm{KL}\left(Q||P\right)\leq\kappa\}$ and $\mathcal{G}_\kappa=\{\mathbb{E}_Q f(\cdot):Q\in\mathcal{Q}(\kappa)\}$. Then for any $g\in\mathcal{G}_\kappa$, there exists a $Q\in\mathcal{Q}(\kappa)$ such that $g=\mathbb{E}_Q f(\cdot)$. We can first show a tail bound that for any given $c_2>0$ and $g\in\mathcal{G}_\kappa$, if $t\geq\frac{(1+c_2)(1+c_2h^2)}{mc_2h^2}$, then

$$\mathbb{P}_S\left(\mathcal{L}_{\mathcal{D}}(g)-(1+c_2)\hat{\mathcal{L}}_S(g)+c_2(1-h^2)\hat{\mathcal{L}}_S(g^2)\geq\frac{t}{2}\right)\leq\frac{1}{2}. \tag{5.17}$$

Then, consider another independent random data set $S' = \{z'_1, \ldots, z'_m\} \in \mathcal{D}^m$. For $c > c_2$, by taking the difference of $\mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2)$ and $\mathcal{L}_{\mathcal{D}}(g) - (1+c_2)\hat{\mathcal{L}}_{S'}(g) + c_2(1-h^2)\hat{\mathcal{L}}_{S'}(g^2)$ and using Eq. (5.17), we obtain

$$\frac{1}{4}\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa} \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \geq t\right) \tag{5.18}$$

$$\leq \frac{1}{2}\mathbb{P}_{S,S'}\left(\sup_{g\in\mathcal{G}_\kappa} (1+c_2)\hat{\mathcal{L}}_{S'}(g) - c_2(1-h^2)\hat{\mathcal{L}}_{S'}(g^2) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \geq \frac{t}{2}\right). \tag{5.19}$$

Now by writing $(1+c_2)\hat{\mathcal{L}}_{S'}(g) - c_2(1-h^2)\hat{\mathcal{L}}_{S'}(g^2) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2)$ as

$$\begin{aligned}(1+\frac{c+c_2}{2})\left(\hat{\mathcal{L}}_{S'}(g) - \hat{\mathcal{L}}_S(g)\right) - \frac{c+c_2}{2}(1-h^2)\left(\hat{\mathcal{L}}_{S'}(g^2) - \hat{\mathcal{L}}_S(g^2)\right)\\ - \frac{c-c_2}{2}\hat{\mathcal{L}}_S\left(g - (1-h^2)g^2\right) - \frac{c-c_2}{2}\hat{\mathcal{L}}_{S'}\left(g - (1-h^2)g^2\right),\end{aligned} \tag{5.20}$$

one can apply the symmetrization argument to get

$$\begin{aligned}&\frac{1}{4}\mathbb{P}_S\left(\sup_{g\in\mathcal{G}_\kappa} \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \geq t\right)\\ &\leq \mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\sup_{g\in\mathcal{G}_\kappa}\left[\frac{1}{m}\sum_{i=1}^m \epsilon_i\left((1+c')\,g(z_i) - c'(1-h^2)g^2(z_i)\right) - c''\hat{\mathcal{L}}_S\left(g-(1-h^2)g^2\right)\right] \geq \frac{t}{4}\right),\end{aligned} \tag{5.21}$$

where $c' = \frac{c+c_2}{2}, c'' = \frac{c-c_2}{2}$. Therefore, we have derived the new shifted symmetrization in deviation involving a quadratic term.

Recalling the definition of $\mathcal{G}_\kappa$, we have

$$\begin{aligned}&\sup_{g\in\mathcal{G}_\kappa}\frac{1}{m}\sum_{i=1}^m \epsilon_i\left((1+c')\,g(z_i) - c'(1-h^2)g(z_i)^2\right) - c''\hat{\mathcal{L}}_S(g-(1-h^2)g^2)\\ &= \sup_{Q\in\mathcal{Q}(\kappa)}\frac{1}{m}\sum_{i=1}^m[\epsilon_i\,(1+c') - c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c']\,(1-h^2)[\mathbb{E}_Q f(z_i)]^2.\end{aligned} \tag{5.22}$$

Note that there are two shifted Rademacher random variables $\epsilon_i\,(1+c') - c''$ and $\epsilon_i c' - c''$, which not only involve a shift term $-c''$ but also scale terms $(1+c')$ and c', respectively. Furthermore, the term $[\mathbb{E}_Q f(z_i)]^2$ cannot be seen as a linear function of Q. This prevents the use of the key argument in [KST08] to formulate an upper bound using Rademacher complexities of constrained linear classes by considering the generalization error as a linear function of Q.

In order to sidestep this obstruction, define $\boldsymbol{\epsilon} := \{\epsilon_i\}_{i=1}^m, \mathbf{z} := \{z_i\}_{i=1}^m$ and suppose $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ achieves the supremum above. (If the supremum cannot be achieved, one can use a carefully chosen sequence of $\{\hat{Q}_i(\boldsymbol{\epsilon}, \mathbf{z})\}$ to prove the same statement as the supremum can be approximated arbitrarily closely.) The following inequality then holds:

$$\begin{aligned}&\sup_{Q\in\mathcal{Q}(\kappa)}\frac{1}{m}\sum_{i=1}^m[\epsilon_i\,(1+c') - c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c']\,(1-h^2)[\mathbb{E}_Q f(z_i)]^2\\ &\leq \sup_{Q\in\mathcal{Q}(\kappa)}\frac{1}{m}\sum_{i=1}^m[\epsilon_i\,(1+c') - c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c']\,(1-h^2)\mathbb{E}_Q f(z_i)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i).\end{aligned} \tag{5.23}$$

To see this, note that, on the one hand, if we plug in $Q = \hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ the inequality is tight; on the other hand, by definition, $Q = \hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ already achieves the supremum of the l.h.s. Note that the r.h.s. can be seen as a linear function of Q, because $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ is a random variable which does not depend on Q.

Let $\epsilon_i'' := \epsilon_i c' - c'' = \epsilon_i \frac{c_1+c_2}{2} - \frac{c_1-c_2}{2}$. Then by keeping the term $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$, one can apply the convex conjugate of relative entropy to get

$$\begin{aligned}
&\mathbb{P}\left[\sup_{Q\in\mathcal{Q}(\kappa)} \mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q) + \frac{c(1-h^2)}{m}\sum_{i=1}^m (\mathbb{E}_Q f(z_i))^2 \geq t\right] \\
&\leq 4\exp\left(\kappa - \frac{\lambda t}{4}\right)\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^m f(z_i)\left[(\epsilon_i+\epsilon_i'') - \epsilon_i''(1-h^2)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)\right]\right)\right]. \qquad (5.24)
\end{aligned}$$

Therefore, the problem turns to bounding the expectation of a function involving shifted Rademacher processes. Although the expectation looks quite complicated since it involves two scaled and shifted Rademacher variables as well as the unknown $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$, fortunately, we are able to show that, for any random variables $Y_i \in [0,1]$, we have

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^m f(z_i)\left[(\epsilon_i+\epsilon_i'') - \epsilon_i''(1-h^2)Y_i\right]\right)\right] \leq 1, \qquad (5.25)$$

if $h \in (0,1], 1 > h^2c > c_2 > 0$ and $0 < \frac{\lambda}{m} < C = \frac{h^2c-c_2}{2(1+h^2c)(1+c_2)}$. This result removes the term $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ by letting $Y_i = \mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)$. Finally, we combine different values of κ by a union bound argument similar to the proof of Proposition 5.2.1 to complete the proof. □

5.3.1 Comparison with Catoni's bound

As we have shown in Proposition 5.2.1, using shifted Rademacher processes, we can match Catoni's fast-rate PAC-Bayesian bound (Theorem 5.1.2) up to constants. We have also presented a new fast-rate PAC-Bayes bound based on "flatness". Although both our bound and Catoni's bound show fast $\mathcal{O}(m^{-1})$ rates of convergence, our bound can exploit flatness in the posterior distribution.

In particular, our PAC-Bayes bound based on flatness (Eq. (5.15)) can be much tighter than Catoni's bound (Eq. (5.6)) when the posterior is chosen to concentrate on a "flat minimum" where $\frac{c}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2$ is very small yet $\hat{\mathcal{L}}_S(Q)$ is nonzero. It can be verified that the "flatness" term $\frac{c}{m}\sum_{i=1}^m \mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2$ in Eq. (5.15) is smaller than the excess empirical risk term $c\hat{\mathcal{L}}_S(Q)$ when $\frac{1-h^2}{m}\sum_{i=1}^m (\mathbb{E}_Q f(z_i))^2$ is greater than 0, which is precisely when the empirical risk is greater than zero. (See Eq. (5.14).)

Based on this observation, we expect our bound to be tighter for sufficient flat posteriors, nonzero empirical risk, and sufficient training data. In order to see this, note that Catoni's bound has the form $(1+c_c)\hat{\mathcal{L}}_S Q + \frac{\mathcal{C}_c}{m}(\mathrm{KL}(Q\|P) + \log\frac{1}{\delta})$, while our bound based on Eq. (5.14) can be written $(1+c_r)\hat{\mathcal{L}}_S Q - \frac{c_r(1-h^2)}{m}\sum_{i=1}^m (\mathbb{E}_Q f(z_i))^2 + \frac{\mathcal{C}_r}{m}(\mathrm{KL}(Q\|P) + \log\frac{1}{\delta} + 1)$. Here c_c, c_r inflate the empirical risk and $\mathcal{C}_c, \mathcal{C}_r$ are constants. Let T_m be $\frac{c_r(1-h^2)}{m}\sum_{i=1}^m (\mathbb{E}_Q f(z_i))^2$. Note that c_c and c_r must be fixed before seeing the data. Assuming we equate the inflation of the empirical risk, i.e., $c_c = c_r$, the proposed bound is tighter than Catoni's bound provided $m > \frac{1}{T_m}\left((\mathcal{C}_r - \mathcal{C}_c)\left(\mathrm{KL}(Q\|P) + \log\frac{1}{\delta}\right) + \mathcal{C}_r\right)$. If T_m converges to a positive number (a reasonable assumption), then our proposed bound will be tighter for sufficiently many samples. If we assume $c_c \neq c_r$, our bound can still be tighter than Catoni's bound under more involved conditions.

5.4 Related work and conclusion

5.4.1 Related work

There is a large literature on obtaining fast $1/m$ convergence rates for generalization error and excess risk using Rademacher processes and their generalizations [BBM05; Kol06; LM12; LRS15; ZH18]. As far as we know, this literature does not connect with the PAC-Bayesian literature. There do exist, however, PAC-Bayesian analyses for specific learning algorithms that achieve fast rates [Aud09; Lac+07; Ger+15]. These specific analyses do not lead to general PAC-Bayes bounds, like those produced by Catoni [Cat07].

Our new PAC-Bayes bound based on flatness bears a superficial resemblance to a number of bounds in the literature. However, our notion of flatness is *not* related to the variance of the randomized classifier caused by the randomness of the observed data. Therefore, our new bound is fundamentally different from existing PAC-Bayes bounds based on this type of variance [Lac+07; Ger+15; TS13].

For example, Tolstikhin and Seldin [TS13, Thm. 4] presents a generalization bound based on the "empirical variance", which is distinct from our "flatness". The "empirical variance" is $\mathbb{E}_Q \frac{1}{m} \sum_{i=1}^m [f(z_i) - \frac{1}{m} \sum_{i=1}^m f(z_i)]^2$, while our "flatness" is $\frac{1}{m} \sum_{i=1}^m \mathbb{E}_Q [f(z_i) - \mathbb{E}_Q f(z_i)]^2$. Note that it is possible for flatness to be zero, even when empirical variance is large.

To the best of our knowledge, the closest work to ours in the literature is that by Audibert [Aud09]. The bound given in [Aud09, Thm. 6.1] uses a notion similar to our "flatness". The bound is, however, not comparable with ours for several reasons: First, [Aud09, Theorem 6.1] holds only for the particular algorithm proposed by Audibert, and so it is not a general PAC-Bayes bound like ours. Second, our notion of "flatness" is empirical, while the "flatness" term in [Aud09, Theorem 6.1] is defined by an expectation over the data distribution, which is often presumed unknown. Finally, the proof techniques used to establish [Aud09, Theorem 6.1] are specialized to the proposed algorithm and not based on the use of Rademacher processes. Our proof techniques via shifted Rademacher processes provides a blueprint for other approaches to deriving fast-rate PAC-Bayes bounds.

Grünwald and Mehta [GM19] establish new excess risk bounds in terms of a novel complexity measure based on "luckiness" functions. In the setting of randomized classifiers, particular choices of luckiness functions can be related to PAC-Bayesian notions of complexity based on "priors". Indeed, in this setting, their complexity measure can be bounded in terms of a KL divergence, as in PAC-Bayesian bounds. In a setting with deterministic classifiers, the authors show that their complexity measure can be bounded in terms of Rademacher complexity. Thus, while their framework connects with both PAC-Bayesian and Rademacher-complexity bounds, it is not immediately clear whether it produces direct connections, as we have accomplished here. It is certainly interesting to consider whether our bounds can be achieved (or surpassed) by an appropriate use of their framework.

5.4.2 Conclusion

In this chapter we exploit the connections between modern PAC-Bayesian theory and Rademacher complexities. Using shifted Rademacher processes [Weg03; LM12; ZH18], we derive a novel fast-rate PAC-Bayes bound that depends on the empirical "flatness" of the posterior. Our work provides new insights on PAC-Bayesian theory and opens up new avenues for developing stronger bounds.

It is worth highlighting some potentially interesting directions that may be worth further investigation:

We have "rederived" Catoni's bound via shifted Rademacher processes, up to constants. It is

interesting to ask whether the Rademacher approach can dominate the direct PAC-Bayes bound. In the other direction, we have not derived our flatness bound via a direct PAC-Bayes approach. Whether this is possible and what it achieves might shed light on the relative strengths of these two distinct approaches to PAC-Bayes bounds. It may also be interesting to pursue PAC-Bayes bounds via some adaptation of Talagrand's concentration inequalities [Wai19, Ch.3].

We have derived PAC-Bayes bounds for zero–one loss. While the extension to bounded loss is straightforward, the problem of extending our approach to unbounded loss relates to a growing body of work on this problem within the PAC-Bayesian framework. (See, for example, [AG18] and the references therein). Whether the Rademacher perspective is helpful or not in this regard is not clear at this point.

There has been a surge of interest in PAC-Bayes bounds and their application to the study of generalization in large-scale neural networks. One promising direction is to consider Rademacher-process techniques may aid in the development of PAC-Bayesian analyses of specific algorithms [Aud09; Lac+07; Ger+15], especially in the case when the algorithms are related to large-scale neural networks trained by stochastic gradient descent [NBS17; Ney+17; Lon17].

It would be interesting to perform a careful empirical study of our flatness bound in the context of large-scale neural networks, in the vein of the work of Dziugaite and Roy [DR17]. Preliminary work suggests that the posteriors found by PAC-Bayes bound optimization are not flat in our sense. After some investigation, we believe the reason is that optimizing the PAC-Bayes bound results in underfitting, due in part to the distribution-independent prior. It would be interesting to compare various PAC-Bayes bounds under strict constraints on the empirical risk.

5.5 Appendices

5.5.1 Proof of Proposition 5.2.1

To match Catoni's bound, we need to control $\sup_{f\in\mathcal{F}} \mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f)$, given $c > 0$. We apply an existing result due to Zhivotovskiy and Hanneke [ZH18], which we quote here:

Lemma 5.5.1 (Shifted symmetrization in deviation [ZH18, Cor. 7]). *Fix constants $c > c_2 > 0$,*

$$\frac{1}{4}\mathbb{P}_S\left(\sup_{f\in\mathcal{F}} \mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f) \geq t\right) \leq \mathbb{P}_{S,\epsilon}\left(\sup_{f\in\mathcal{F}}\left[\frac{1+c'/2}{m}\sum_{i=1}^{m}\left(\epsilon_i - \frac{c'/2}{1+c'/2}\right)f(z_i)\right] \geq \frac{t'}{2}\right). \tag{5.26}$$

where $c' = \frac{c-c_2}{1+c_2}, t' = \frac{t}{2(1+c_2)}$, and $\{\epsilon_i\}$ are Rademacher random variables.

Let $\epsilon_i' := \epsilon_i - \frac{c'}{2+c'}$. For $\kappa > 0$, define $\mathcal{Q}(\kappa) := \{Q : \mathrm{KL}\,(Q||P) \leq \kappa\}$. Next, using convex conjugate of the Kullback–Leibler divergence (the change-measure inequality), for $\lambda > 0$,

$$\sup_{Q\in\mathcal{Q}(\kappa)}\left[\frac{1}{m}\sum_{i=1}^{m}\epsilon_i'\mathbb{E}_Q f(z_i) - \frac{1}{\lambda}\mathrm{KL}\,(Q||P)\right] \leq \frac{1}{\lambda}\log\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right], \tag{5.27}$$

which implies

$$\sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m} \epsilon_i' \mathbb{E}_Q f(z_i) \le \frac{\kappa}{\lambda} + \frac{1}{\lambda}\log \mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right]. \tag{5.28}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{P}_S\left(\sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f)] \ge t\right) && (5.29)\\
&\quad= \mathbb{P}_S\left(\sup_{Q\in\mathcal{Q}(\kappa)} \mathcal{L}_{\mathcal{D}}(\mathbb{E}_Q f) - (1+c)\hat{\mathcal{L}}_S(\mathbb{E}_Q f) \ge t\right) && (5.30)\\
&\quad\le 4\mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}(\epsilon_i - \frac{c'}{2+c'})\mathbb{E}_Q f(z_i) \ge \frac{t'}{2+c'}\right) && (5.31)\\
&\quad\le 4\mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\frac{\kappa}{\lambda} + \frac{1}{\lambda}\log \mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right] \ge \frac{t'}{2+c'}\right) && (5.32)\\
&\quad= 4\mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\log \mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right] \ge \frac{\lambda t'}{2+c'} - \kappa\right) && (5.33)\\
&\quad= 4\mathbb{P}_{S,\boldsymbol{\epsilon}}\left(\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right] \ge \exp\left(\frac{\lambda t'}{2+c'} - \kappa\right)\right) && (5.34)\\
&\quad\underbrace{\le}_{\text{Markov}} 4\exp\left(\kappa - \frac{\lambda t'}{2+c'}\right)\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}\epsilon_i' f(z_i)\right)\right]. && (5.35)
\end{aligned}$$

Now we use the result of Lemma 5.5.2. For any c_2 such that $0 < c_2 < c$, if $t \ge \frac{1}{m}\frac{(1+c_2)^2}{c_2}$ and $\frac{c'}{c'+2} \ge \frac{\log\cosh(\lambda/m)}{\lambda/m}$ then we have

$$\mathbb{P}\left(\sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f)] \ge t\right) \le 4\exp\left(\kappa - \frac{\lambda t'}{2+c'}\right), \tag{5.36}$$

where $c' = \frac{c-c_2}{1+c_2}$ and $t' = \frac{t}{2(1+c_2)}$. Now letting $4\exp\left(\kappa - \frac{\lambda t'}{2+c'}\right)$ equals to δ, we have

$$t = 2(1+c_2)t' = 2(1+c_2)\left[\frac{2+c'}{\lambda}\left(\kappa + \log(4/\delta)\right)\right]. \tag{5.37}$$

Now let $\frac{\lambda}{m} = C$, noting that

$$t \ge 2(1+c_2)\left[\frac{2+c'}{\lambda}\log(4/\delta)\right] \tag{5.38}$$

then we can choose C small enough (clearly bounded away from 0) to satisfy

$$\frac{\log\cosh(C)}{C} \le \frac{c'}{c'+2}, \quad C \le \frac{2(1+c_2)(2+c')\log(4/\delta)}{(1+c_2)^2/c_2} \tag{5.39}$$

which guarantees both $t \ge \frac{1}{m}\frac{(1+c_2)^2}{c_2}$ and $\frac{c'}{c'+2} \ge \frac{\log\cosh(\lambda/m)}{\lambda/m}$.

Therefore, using such C we have

$$\mathbb{P}\left(\sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f)] \geq \frac{2(1+c_2)(2+c')}{Cm}(\kappa + \log(4/\delta))\right) \leq \delta, \tag{5.40}$$

which implies

$$\sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q[\mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f)] \leq \frac{C'}{m}(\kappa + \log(4/\delta)), \quad \text{w.p.} \quad 1-\delta. \tag{5.41}$$

where $C' = \frac{2(1+c_2)(2+c')}{C}$.

Finally, we combine all possible κ using a union bound. Define $\Gamma_0 := \{Q : \mathrm{KL}\,(Q||P) \leq 2\}$ and $\Gamma_j := \{Q : \mathrm{KL}\,(Q||P) \in [2^j, 2^{j+1}]\}$ for $j \geq 1$. Let $\delta_j = 2^{-(j+1)}\delta$ so $\sum_{j=0}^{\infty}\delta_j = \delta$. Then for any Q there is a j_Q such that $Q \in \Gamma_{j_Q}$. Then by definition we have

$$\mathrm{KL}\,(Q||P) \leq 2^{j_Q+1} \leq 2\max\{\mathrm{KL}\,(Q||P), 1\} \tag{5.42}$$

$$\delta_{j_Q} = 2^{-(j_Q+1)}\delta \geq \frac{\delta}{2\max\{\mathrm{KL}\,(Q||P), 1\}}. \tag{5.43}$$

Therefore, we have that for any Q, with probability $1-\delta$, we have

$$\mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q) \leq \frac{C'}{m}\left(2\max\{\mathrm{KL}\,(Q||P), 1\} + \log(8\max\{\mathrm{KL}\,(Q||P), 1\}/\delta)\right) \tag{5.44}$$

$$\leq \frac{C'}{m}\left(2\max\{\mathrm{KL}\,(Q||P), 1\} + \log(\max\{\mathrm{KL}\,(Q||P), 1\}) + \log(8/\delta)\right). \tag{5.45}$$

Now we simplify the order without optimizing the constants, which gives

$$\log(\max\{\mathrm{KL}\,(Q||P), 1\}) \leq \max\{\mathrm{KL}\,(Q||P), 1\} \leq \mathrm{KL}\,(Q||P) + 1. \tag{5.46}$$

Therefore, we have

$$\mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f) \leq \frac{C_1}{m}\mathrm{KL}\,(Q||P) + \frac{C_2}{m}\log(1/\delta) + \frac{C_3}{m}, \quad \text{w.p.} \quad 1-\delta. \tag{5.47}$$

Lemma 5.5.2. *If $k \geq \frac{\log\cosh(\lambda/m)}{\lambda/m}$, then*

$$\mathbb{E}_P\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m}(\epsilon_i - k)f(z_i)\right)\right] \leq 1. \tag{5.48}$$

Proof. Noting that $\{f(z_i)\}$ are independent Bernoulli random variables, we have

$$\mathbb{E}_P\mathbb{E}_S\mathbb{E}_\epsilon\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^m(\epsilon_i-k)f(z_i)\right)\right] \tag{5.49}$$

$$\underbrace{=}_{\text{indep}}\mathbb{E}_P\prod_{i=1}^m\mathbb{E}_S\mathbb{E}_{\epsilon_i}\left[\exp\left(\frac{\lambda}{m}(\epsilon_i-k)f(z_i)\right)\right] \tag{5.50}$$

$$\underbrace{=}_{\text{Rademacher}}\mathbb{E}_P\prod_{i=1}^m\mathbb{E}_S\left[\frac{e^{(1-k)\frac{\lambda}{m}f(z_i)}+e^{-(1+k)\frac{\lambda}{m}f(z_i)}}{2}\right] \tag{5.51}$$

$$\underbrace{=}_{\text{Bernoulli}}\mathbb{E}_P\prod_{i=1}^m\left[(1-\mathbb{E}_S[f(z_i)])+\left(\frac{e^{\frac{\lambda}{m}}+e^{-\frac{\lambda}{m}}}{2e^{k\frac{\lambda}{m}}}\right)\mathbb{E}_S[f(z_i)]\right] \tag{5.52}$$

$$=\mathbb{E}_P\prod_{i=1}^m\left[(1-\mathbb{E}_S[f(z_i)])+\frac{\cosh\left(\frac{\lambda}{m}\right)}{e^{k\frac{\lambda}{m}}}\mathbb{E}_S[f(z_i)]\right], \tag{5.53}$$

which is upper bounded by 1 if we choose k such that

$$\cosh(\lambda/m)=\frac{e^{\lambda/m}+e^{-\lambda/m}}{2}\le e^{k\lambda/m}. \tag{5.54}$$

That is

$$k\ge\frac{\log\cosh(\lambda/m)}{\lambda/m}. \tag{5.55}$$

□

5.5.2 Proof of Theorem 5.3.3

We first present some lemmas that will be used in the later proof.

Lemma 5.5.3 (Shifted-flatness inequality)**.** *Consider a function $f:\mathcal{Z}\to[0,1]$, constants $h\in[0,1]$ and $c_2>0$, if $t\ge\frac{(1+c_2)(1+c_2h^2)}{mc_2h^2}$, we have*

$$\mathbb{P}_S\left(\mathcal{L}_\mathcal{D}(f)-(1+c_2)\hat{\mathcal{L}}_S(f)+c_2(1-h^2)\hat{\mathcal{L}}_S(f^2)\ge\frac{t}{2}\right)\le\frac{1}{2}. \tag{5.56}$$

Proof. Let $v=c_2\mathcal{L}_\mathcal{D}(f)-c_2(1-h^2)\mathcal{L}_\mathcal{D}(f^2)=c_2\mathcal{L}_\mathcal{D}(f-(1-h^2)f^2)$. Then, we have

$$\mathbb{P}\left(\mathcal{L}_\mathcal{D}(f)-(1+c_2)\hat{\mathcal{L}}_S(f)+c_2(1-h^2)\hat{\mathcal{L}}_S(f^2)\ge\frac{t}{2}\right) \tag{5.57}$$

$$=\mathbb{P}\left(\mathcal{L}_\mathcal{D}(f)-\frac{c_2(1-h^2)}{1+c_2}\mathcal{L}_\mathcal{D}(f^2)-\hat{\mathcal{L}}_S(f)+\frac{c_2(1-h^2)}{1+c_2}\hat{\mathcal{L}}_S(f^2)\ge\frac{t/2+v}{1+c_2}\right) \tag{5.58}$$

$$=\mathbb{P}\left(\mathbb{E}_{z\sim\mathcal{D}}(f(z)-\frac{c_2(1-h^2)}{1+c_2}f(z)^2)-\frac{1}{m}\sum_{i=1}^m(f(z_i)-\frac{c_2(1-h^2)}{1+c_2}f(z_i)^2)\ge\frac{t/2+v}{1+c_2}\right). \tag{5.59}$$

Because $f(z_i)-\frac{c_2(1-h^2)}{1+c_2}f(z_i)^2, i=1,\dots,m$ are i.i.d. random samples, using Chebyshev's inequality

together with $4ab \le (a+b)^2$ and $f \in [0,1]$, the formula above is upper bounded by

$$\frac{(1+c_2)^2 \operatorname{var}(f - \frac{c_2(1-h^2)}{1+c_2} f^2)}{m(t/2+v)^2} \le \frac{(1+c_2)^2 \mathcal{L}_{\mathcal{D}}(f - \frac{c_2(1-h^2)}{1+c_2} f^2)^2}{2mvt} \le \frac{(1+c_2)^2 \mathcal{L}_{\mathcal{D}}(f - \frac{c_2(1-h^2)}{1+c_2} f^2)}{2mvt}. \quad (5.60)$$

We can further decompose the term in the numerator by

$$\mathcal{L}_{\mathcal{D}}(f - \frac{c_2(1-h^2)}{1+c_2} f^2) = \frac{c_2}{1+c_2} \mathcal{L}_{\mathcal{D}}(f - (1-h^2)f^2) + \frac{1}{1+c_2} \mathcal{L}_{\mathcal{D}}(f) \quad (5.61)$$

$$\le \frac{1}{1+c_2} v + \frac{1}{1+c_2} \frac{1}{c_2 h^2} v = \frac{c_2 h^2 + 1}{(1+c_2) c_2 h^2} v, \quad (5.62)$$

Therefore the lemma follows directly from $t \ge \frac{(1+c_2)(1+c_2h^2)}{mc_2h^2}$. □

Lemma 5.5.4 (New shifted symmetrization in deviation)**.** *Fix constants $c > c_2 > 0$, $h \in [0,1]$, if $t \ge \frac{(1+c_2)(1+c_2h^2)}{mc_2h^2}$, we have*

$$\frac{1}{4}\mathbb{P}_S \left(\sup_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f) + c(1-h^2)\hat{\mathcal{L}}_S(f^2) \ge t \right)$$

$$\le \mathbb{P}_{S,\boldsymbol{\epsilon}} \left(\sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_{i=1}^m \epsilon_i \left((1+c') f(z_i) - c'(1-h^2) f^2(z_i) \right) - c'' \hat{\mathcal{L}}_S \left(f - (1-h^2) f^2 \right) \right] \ge \frac{t}{4} \right), \quad (5.63)$$

where $c' = \frac{c+c_2}{2}, c'' = \frac{c-c_2}{2}$, $\boldsymbol{\epsilon} := \{\epsilon_i\}_{i=1}^m$, in which $\{\epsilon_i\}$ are independent Rademacher random variables.

Proof. Consider a random set $S' = \{z_i'\}_{i=1}^m \in \mathcal{D}^m$, in Lemma 5.5.3 we have shown that

$$\mathbb{P}_{S'} \left(\mathcal{L}_{\mathcal{D}}(f) - (1+c_2)\hat{\mathcal{L}}_{S'}(f) + c_2(1-h^2)\hat{\mathcal{L}}_{S'}(f^2) \ge \frac{t}{2} \right) \le \frac{1}{2}. \quad (5.64)$$

Therefore, we can get

$$\frac{1}{4}\mathbb{P}_S \left(\sup_{f \in \mathcal{F}} \mathcal{L}_{\mathcal{D}}(f) - (1+c)\hat{\mathcal{L}}_S(f) + c(1-h^2)\hat{\mathcal{L}}_S(f^2) \ge t \right) \quad (5.65)$$

$$\le \frac{1}{2}\mathbb{P}_{S,S'} \left(\sup_{f \in \mathcal{F}} (1+c_2)\hat{\mathcal{L}}_{S'}(f) - c_2(1-h^2)\hat{\mathcal{L}}_{S'}(f^2) - (1+c)\hat{\mathcal{L}}_S(f) + c(1-h^2)\hat{\mathcal{L}}_S(f^2) \ge \frac{t}{2} \right) \quad (5.66)$$

$$= \frac{1}{2}\mathbb{P}_{S,S'} \left[\sup_{f \in \mathcal{F}} (1 + \frac{c+c_2}{2}) \left(\hat{\mathcal{L}}_S(f) - \hat{\mathcal{L}}_{S'}(f) \right) - \frac{c+c_2}{2}(1-h^2) \left(\hat{\mathcal{L}}_S(f^2) - \hat{\mathcal{L}}_{S'}(f^2) \right) \right. \quad (5.67)$$

$$\left. - \frac{c-c_2}{2} \hat{\mathcal{L}}_{S'} \left(f - (1-h^2)f^2 \right) - \frac{c-c_2}{2} \hat{\mathcal{L}}_S \left(f - (1-h^2)f^2 \right) \ge \frac{t}{2} \right] \quad (5.68)$$

$$\le \mathbb{P}_{S,\boldsymbol{\epsilon}} \left(\sup_{f \in \mathcal{F}} \left[\frac{1}{m} \sum_{i=1}^m \epsilon_i \left((1+c') f(z_i) - c'(1-h^2) f^2(z_i) \right) - c'' \hat{\mathcal{L}}_S \left(f - (1-h^2) f^2 \right) \right] \ge \frac{t}{4} \right), \quad (5.69)$$

where $c' = \frac{c+c_2}{2}, c'' = \frac{c-c_2}{2}$, and the last inequality is by the symmetrization argument. □

Lemma 5.5.5. *For constants $h \in (0,1], h^2c > c_2 > 0$, let $C = \frac{h^2c - c_2}{2(1+h^2c)(1+c_2)}$, if $0 < \frac{\lambda}{m} < C$, then given*

independent Bernoulli random variables X_i, for any random variables $Y_i \in [0,1]$,

$$\mathbb{E}_\epsilon \mathbb{E}_X \mathbb{E}_{Y|X,\epsilon} \left[\exp\left(\frac{\lambda}{m} \sum_{i=1}^{m} X_i \left[(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) Y_i \right] \right) \right] \leq 1, \tag{5.70}$$

where $\{\epsilon_i\}$ are i.i.d. Rademacher random variables and $\epsilon_i'' = \epsilon_i \frac{c+c_2}{2} - \frac{c-c_2}{2}$.

Proof. Note when $X_i = 0$, the value of Y_i has no effect onto LHS. When $X_i = 1$,

$$(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) Y_i = (1+c_2) - c_2(1-h^2) Y_i, \text{ if } \epsilon_i = 1, \tag{5.71}$$

$$(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) Y_i = -(1+c) + c(1-h^2) Y_i, \text{ if } \epsilon_i = -1, \tag{5.72}$$

Therefore, we have

$$(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) Y_i \leq (\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) \frac{1-\epsilon_i}{2}. \tag{5.73}$$

Denoting $\mu_i = \mathbb{E}[X_i]$, by the monotonicity of the exponential function, we have

$$\mathbb{E}_\epsilon \mathbb{E}_X \mathbb{E}_{Y|X,\epsilon} \left[\exp\left(\frac{\lambda}{m} \sum_{i=1}^{m} X_i \left[(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) Y_i \right] \right) \right] \tag{5.74}$$

$$\leq \mathbb{E}_\epsilon \mathbb{E}_X \left[\exp\left(\frac{\lambda}{m} \sum_{i=1}^{m} X_i \left[(\epsilon_i + \epsilon_i'') - \epsilon_i''(1-h^2) \frac{1-\epsilon_i}{2} \right] \right) \right] \tag{5.75}$$

$$= \prod_{i=1}^{m} \left[1 - \mu_i + \frac{\mu_i}{2} \exp\left(\frac{\lambda}{m}(1+c_2) \right) + \frac{\mu_i}{2} \exp\left(-\frac{\lambda}{m}(1+h^2 c) \right) \right], \tag{5.76}$$

For the formula upper bounded by 1, it is sufficient to prove

$$\exp\left(\frac{\lambda}{m} (1+c_2) \right) + \exp\left(-\frac{\lambda}{m} \left(1 + h^2 c\right) \right) \leq 2. \tag{5.77}$$

Because we have $e^x \geq x+1$, thus $e^{-x} \leq \frac{1}{1+x}$ for $x > -1$ and $e^x \leq \frac{1}{1-x}$ for $x < 1$. Therefore, it is sufficient to have

$$\frac{1}{1 - \frac{\lambda}{m}(1+c_2)} + \frac{1}{1 + \frac{\lambda}{m}(1+h^2 c)} \leq 2, \quad \frac{\lambda}{m}(1+c_2) \leq 1. \tag{5.78}$$

Thus, we know the argument holds for

$$\frac{\lambda}{m} \leq \min\left\{ \frac{h^2 c - c_2}{2(1+h^2 c)(1+c_2)}, \frac{1}{1+c_2} \right\} = \frac{h^2 c - c_2}{2(1+h^2 c)(1+c_2)}, \tag{5.79}$$

where the last equality holds when $c_2 < h^2 c$. □

Now we are ready for the proof of Theorem 5.3.3. Denoting $g(\cdot) = \mathbb{E}_Q f(\cdot)$, one can write

$$\mathcal{L}_{\mathcal{D}}(Q) - \hat{\mathcal{L}}_S(Q) - \frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2 \tag{5.80}$$

$$= \mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q) + \frac{c(1-h^2)}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2 \tag{5.81}$$

$$= \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + \frac{c(1-h^2)}{m}\sum_{i=1}^{m}(g(z_i))^2 \tag{5.82}$$

$$= \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2). \tag{5.83}$$

Recall that for $\kappa > 0$, we have defined $\mathcal{Q}(\kappa) = \{Q : \mathrm{KL}\,(Q||P) \le \kappa\}$. We start from the formula,

$$\mathbb{P}_S\left[\sup_{Q\in\mathcal{Q}(\kappa)} \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \ge t\right]. \tag{5.84}$$

By Lemma 5.5.4, let $c' = \frac{c+c_2}{2}, c'' = \frac{c-c_2}{2}$, and $\{\epsilon_i\}$ being i.i.d. Rademacher random variables, we have

$$\mathbb{P}_S\left[\sup_{Q\in\mathcal{Q}(\kappa)} \mathcal{L}_{\mathcal{D}}(g) - (1+c)\hat{\mathcal{L}}_S(g) + c(1-h^2)\hat{\mathcal{L}}_S(g^2) \ge t\right] \tag{5.85}$$

$$\le 4\mathbb{P}_{S,\epsilon}\left[\sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\left((1+c')\,g(z_i) - c'(1-h^2)g(z_i)^2\right) - c''\hat{\mathcal{L}}_S(g-(1-h^2)g^2) \ge \frac{t}{4}\right]. \tag{5.86}$$

Plugging into $g = \mathbb{E}_Q f(\cdot)$ yields

$$\sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}\epsilon_i\left((1+c')\,g(z_i) - c'(1-h^2)g(z_i)^2\right) - c''\hat{\mathcal{L}}_S(g-(1-h^2)g^2) \tag{5.87}$$

$$= \sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}[\epsilon_i\,(1+c') - c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c'']\,(1-h^2)[\mathbb{E}_Q f(z_i)]^2. \tag{5.88}$$

Given $\boldsymbol{\epsilon} = \{\epsilon_i\}_{i=1}^m, \mathbf{z} = \{z_i\}_{i=1}^m$, we suppose $\hat{Q}(\boldsymbol{\epsilon}, \mathbf{z})$ achieves the supremum above (if the supremum cannot be achieved, one can use a sequence of $\{\hat{Q}_i(\boldsymbol{\epsilon}, \mathbf{z})\}$ to approximate arbitrarily close to the supremum). Using

$$\epsilon_i'' := \epsilon_i c' - c'' = \epsilon_i\frac{c_1+c_2}{2} - \frac{c_1-c_2}{2}, \tag{5.89}$$

we have

$$\sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}[\epsilon_i(1+c')-c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c'']\,(1-h^2)[\mathbb{E}_Q f(z_i)]^2 \tag{5.90}$$

$$\leq \sup_{Q\in\mathcal{Q}(\kappa)} \frac{1}{m}\sum_{i=1}^{m}[\epsilon_i(1+c')-c'']\mathbb{E}_Q f(z_i) - [\epsilon_i c' - c'']\,(1-h^2)\mathbb{E}_Q f(z_i)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i) \tag{5.91}$$

$$= \sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q\left[\frac{1}{m}\sum_{i=1}^{m}(\epsilon_i+\epsilon_i'')f(z_i) - \epsilon_i''(1-h^2)f(z_i)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)\right] \tag{5.92}$$

$$\leq \frac{\kappa}{\lambda} + \frac{1}{\lambda}\log\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} f(z_i)\left[(\epsilon_i+\epsilon_i'') - \epsilon_i''(1-h^2)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)\right]\right)\right], \tag{5.93}$$

where the last inequality follows from duality for convex optimization [BV04, Chp. 5].

Therefore, we have

$$\mathbb{P}\left[\sup_{Q\in\mathcal{Q}(\kappa)} \mathcal{L}_{\mathcal{D}}(Q) - (1+c)\hat{\mathcal{L}}_S(Q) + \frac{c(1-h^2)}{m}\sum_{i=1}^{m}(\mathbb{E}_Q f(z_i))^2 \geq t\right] \tag{5.94}$$

$$\leq 4\mathbb{P}\left[\frac{\kappa}{\lambda} + \frac{1}{\lambda}\log\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} f(z_i)\left[(\epsilon_i+\epsilon_i'') - \epsilon_i''(1-h^2)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)\right]\right)\right] \geq \frac{t}{4}\right] \tag{5.95}$$

$$\leq 4\exp\left(\kappa - \frac{\lambda t}{4}\right)\mathbb{E}_S\mathbb{E}_{\boldsymbol{\epsilon}}\mathbb{E}_P\left[\exp\left(\frac{\lambda}{m}\sum_{i=1}^{m} f(z_i)\left[(\epsilon_i+\epsilon_i'') - \epsilon_i''(1-h^2)\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)\right]\right)\right] \tag{5.96}$$

$$\leq 4\exp\left(\kappa - \frac{\lambda t}{4}\right), \tag{5.97}$$

where the last inequality comes from Lemma 5.5.5 by considering X_i as $f(z_i)$ and Y_i as $\mathbb{E}_{\hat{Q}(\boldsymbol{\epsilon},\mathbf{z})} f(z_i)$, with

$$C := \frac{\lambda}{m} \leq \frac{h^2c - c_2}{2(1+h^2c)(1+c_2)}. \tag{5.98}$$

Now let $4\exp\left(\kappa - \frac{\lambda t}{4}\right)$ equals to δ, we have

$$t = \frac{4}{\lambda}(\kappa + \log\frac{4}{\delta}) = \frac{4}{Cm}(\kappa + \log\frac{4}{\delta}). \tag{5.99}$$

Note that the shifted symmetrization inequality requires $t \geq \frac{(1+c_2)(1+c_2h^2)}{mc_2h^2}$ by Lemma 5.5.3. Combining with the previous requirement for C together, we have

$$C \leq \min\left\{\frac{h^2c - c_2}{2(1+h^2c)(1+c_2)}, \frac{4c_2h^2}{(1+c_2)(1+c_2h^2)}(\kappa + \log\frac{4}{\delta})\right\}. \tag{5.100}$$

Using such C we have with probability at least $1-\delta$,

$$\sup_{Q\in\mathcal{Q}(\kappa)} \mathbb{E}_Q\mathcal{L}_{\mathcal{D}}(f) - \mathbb{E}_Q\hat{\mathcal{L}}_S(f) - \frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i) - (1+h)\mathbb{E}_Q f(z_i)]^2 \leq \frac{4}{Cm}(\kappa + \log\frac{4}{\delta}). \tag{5.101}$$

Finally we combine all possible κ using a union bound. Define $\Gamma_0 = \{Q : \mathrm{KL}(Q||P) \leq 2\}$ and $\Gamma_j = \{Q : \mathrm{KL}(Q||P) \in [2^j, 2^{j+1}]\}$ for $j \geq 1$. Let $\delta_j = 2^{-(j+1)}\delta$ so that $\sum_{j=0}^{\infty}\delta_j = \delta$. Then for any Q

there is a j_Q such that $Q \in \Gamma_{j_Q}$. Then we have

$$\mathrm{KL}\left(Q||P\right) \leq 2^{j_Q+1} \leq 2\max(\mathrm{KL}\left(Q||P\right),1) \tag{5.102}$$

$$\delta_{j_Q} = 2^{-(j_Q+1)}\delta \geq \frac{\delta}{2\max(\mathrm{KL}\left(Q||P\right),1)}, \tag{5.103}$$

Therefore with probability at least $1-\delta$ over draws of S, for any Q,

$$\begin{aligned}\mathbb{E}_Q\mathcal{L}_{\mathcal{D}}(f) &\leq \mathbb{E}_Q\hat{\mathcal{L}}_S(f) + \frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2 \\ &\quad + \frac{4}{Cm}\left[2\max(\mathrm{KL}\left(Q||P\right),1) + \log\frac{8\max(\mathrm{KL}\left(Q||P\right),1)}{\delta}\right] \qquad (5.104)\\ &\leq \mathbb{E}_Q\hat{\mathcal{L}}_S(f) + \frac{c}{m}\sum_{i=1}^{m}\mathbb{E}_Q[f(z_i)-(1+h)\mathbb{E}_Q f(z_i)]^2 + \frac{4}{Cm}\left[3\mathrm{KL}\left(Q||P\right) + \log\frac{1}{\delta}+5\right], \qquad (5.105)\end{aligned}$$

provided that

$$C \leq \min\left\{\frac{h^2c-c_2}{2(1+h^2c)(1+c_2)}, \frac{4c_2h^2}{(1+c_2)(1+c_2h^2)}(\kappa+\log\frac{4}{\delta})\right\}. \tag{5.106}$$

Therefore, it is sufficient if

$$C = \frac{2h^4c}{1+16h^2c},\ c_2 = \frac{h^2c}{1+16h^2c}. \tag{5.107}$$

Bibliography

[AB07] J.-Y. Audibert and O. Bousquet. "Combining PAC-Bayesian and Generic Chaining Bounds". *Journal of Machine Learning Research* 8 (2007), pp. 863–889.

[AG18] P. Alquier and B. Guedj. "Simpler PAC-Bayesian bounds for hostile data". *Machine Learning* 107.5 (2018), pp. 887–902.

[Ait75] J. Aitchison. "Goodness of prediction fit". *Biometrika* 62.3 (1975), pp. 547–554.

[AN00] S.-i. Amari and H. Nagaoka. "Methods of information geometry, volume 191 of Translations of Mathematical Monographs". *American Mathematical Society* (2000), p. 13.

[ARR11] Y. F. Atchadé, G. O. Roberts, and J. S. Rosenthal. "Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo". *Statistics and Computing* 21.4 (2011), pp. 555–568.

[AT08] C. Andrieu and J. Thoms. "A tutorial on adaptive MCMC". *Statistics and Computing* 18.4 (2008), pp. 343–373.

[Aud09] J.-Y. Audibert. "Fast learning rates in statistical inference through aggregation". *The Annals of Statistics* 37.4 (2009), pp. 1591–1646.

[Bax05] P. H. Baxendale. "Renewal theory and computable convergence rates for geometrically ergodic Markov chains". *The Annals of Applied Probability* 15.1B (2005), pp. 700–738.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. "Local Rademacher Complexities". *The Annals of Statistics* 33.4 (2005), pp. 1497–1537.

[BBS09] J. O. Berger, J. M. Bernardo, and D. Sun. "The formal definition of reference priors". *The Annals of Statistics* 37.2 (2009), pp. 905–938. ISSN: 0090-5364.

[BDM12] M. Bédard, R. Douc, and E. Moulines. "Scaling analysis of multiple-try MCMC methods". *Stochastic Processes and their Applications* 122.3 (2012), pp. 758–786.

[Ber66] R. H. Berk. "Limiting behavior of posterior distributions when the model is incorrect". *The Annals of Mathematical Statistics* 37.1 (1966), pp. 51–58.

[Ber70] R. H. Berk. "Consistency a posteriori". *The Annals of Mathematical Statistics* (1970), pp. 894–906.

[Ber79] J.-M. Bernardo. "Reference posterior distributions for Bayesian inference". *Journal of the Royal Statistical Society. Series B. Methodological* 41.2 (1979). With discussion, pp. 113–147. ISSN: 0035-9246.

[Bes+08] A. Beskos, G. Roberts, A. Stuart, and J. Voss. "MCMC methods for diffusion bridges". *Stochastics and Dynamics* 8.03 (2008), pp. 319–350.

[Bes+13] A. Beskos, N. Pillai, G. Roberts, J.-M. Sanz-Serna, and A. Stuart. "Optimal tuning of the hybrid Monte Carlo algorithm". *Bernoulli* 19.5A (2013), pp. 1501–1534.

[BGX08] L. D. Brown, E. I. George, and X. Xu. "Admissible predictive density estimation". *The Annals of Statistics* (2008), pp. 1156–1170.

[BH82] L. D. Brown and J. T. Hwang. "A unified admissibility proof". In: *Statistical decision theory and related topics, III, Vol. 1 (West Lafayette, Ind., 1981)*. Academic Press, New York-London, 1982, pp. 205–230.

[BHW16] P. G. Bissiri, C. C. Holmes, and S. G. Walker. "A general framework for updating belief distributions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.5 (2016), pp. 1103–1130.

[BM02] P. L. Bartlett and S. Mendelson. "Rademacher and Gaussian complexities: Risk bounds and structural results". *Journal of Machine Learning Research* 3 (2002), pp. 463–482.

[BM98] O. Bunke and X. Milhaud. "Asymptotic behavior of Bayes estimates under possibly incorrect models". *The Annals of Statistics* 26.2 (1998), pp. 617–644.

[BPS04] L. A. Breyer, M. Piccioni, and S. Scarlatti. "Optimal scaling of MALA for nonlinear regression". *The Annals of Applied Probability* 14.3 (2004), pp. 1479–1505.

[BR00] L. A. Breyer and G. O. Roberts. "From Metropolis to diffusions: Gibbs states and optimal scaling". *Stochastic Processes and their Applications* 90.2 (2000), pp. 181–206.

[BR08] M. Bédard and J. S. Rosenthal. "Optimal scaling of Metropolis algorithms: Heading toward general target distributions". *Canadian Journal of Statistics* 36.4 (2008), pp. 483–503.

[BREZ18] N. Bou-Rabee, A. Eberle, and R. Zimmer. "Coupling and Convergence for Hamiltonian Monte Carlo" (May 1, 2018). arXiv: http://arxiv.org/abs/1805.00452v2 [math.PR].

[BRH13] N. Bou-Rabee and M. Hairer. "Nonasymptotic mixing of the MALA Algorithm". *IMA Journal of Numerical Analysis* 33.1 (2013), pp. 80–110.

[Bro+11] S. Brooks, A. Gelman, G. Jones, and X.-L. Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

[Bro71] L. D. Brown. "Admissible estimators, recurrent diffusions, and insoluble boundary value problems". *Annals of Mathematical Statistics* 42 (1971), pp. 855–903. ISSN: 0003-4851.

[BRS09] A. Beskos, G. Roberts, and A. Stuart. "Optimal scalings for local Metropolis–Hastings chains on nonproduct targets in high dimensions". *The Annals of Applied Probability* 19.3 (2009), pp. 863–898.

[BS94] J. Bernardo and A. Smith. *Bayesian Theory*. Wiley, 1994.

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Bv15] P. Bühlmann and S. van de Geer. "High-dimensional inference in misspecified linear models". *Electronic Journal of Statistics* 9.1 (2015), pp. 1449–1473.

[Béd07] M. Bédard. "Weak convergence of Metropolis algorithms for non-i.i.d. target distributions". *The Annals of Applied Probability* 17.4 (Aug. 2007), pp. 1222–1244.

[Béd08] M. Bédard. "Optimal acceptance rates for Metropolis algorithms: moving beyond 0.234". *Stochastic Processes and their Applications* 118.12 (2008), pp. 2198–2222.

[Béd19] M. Bédard. "Hierarchical Models and Tuning of Random Walk Metropolis Algorithms". *Journal of Probability and Statistics* 2019 (2019).

[Bég+16] L. Bégin, P. Germain, F. Laviolette, and J.-F. Roy. "PAC-Bayesian bounds based on the Rényi divergence". In: *Artificial Intelligence and Statistics*. 2016, pp. 435–444.

[Cat07] O. Catoni. *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*. Vol. 56. Lecture Notes – Monograph Series. Institute of Mathematical Statistics, 2007.

[CH13] H. M. Choi and J. P. Hobert. "The Polya-Gamma Gibbs sampler for Bayesian logistic regression is uniformly ergodic". *Electronic Journal of Statistics* 7 (2013), pp. 2054–2064.

[Che+18] V. Chen, M. M. Dunlop, O. Papaspiliopoulos, and A. M. Stuart. "Dimension-Robust MCMC in Bayesian Inverse Problems" (Mar. 9, 2018).

[Che85] C.-F. Chen. "On asymptotic normality of limiting density functions with Bayesian implications". *Journal of the Royal Statistical Society. Series B (Methodological)* (1985), pp. 540–546.

[Cob65] A. Cobham. "The Intrinsic Computational Difficulty of Functions". In: *Logic, Methodology and Philosophy of Science: Proceedings of the 1964 International Congress (Studies in Logic and the Foundations of Mathematics)*. Ed. by Y. Bar-Hillel. North-Holland Publishing, 1965, pp. 24–30.

[Coo71] S. A. Cook. "The complexity of theorem-proving procedures". In: *Proceedings of the third annual ACM symposium on Theory of computing*. ACM. 1971, pp. 151–158.

[Cot+13] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. "MCMC methods for functions: modifying old algorithms to make them faster". *Statistical Science* (2013), pp. 424–446.

[CRR05] O. F. Christensen, G. O. Roberts, and J. S. Rosenthal. "Scaling limits for the transient phase of local Metropolis–Hastings algorithms". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2 (2005), pp. 253–268.

[Dal17] A. S. Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.3 (2017), pp. 651–676.

[DF03] M. Dyer and A. Frieze. "Randomly coloring graphs with lower bounds on girth and maximum degree". *Random Structures & Algorithms* 23.2 (2003), pp. 167–179.

[DK19] A. S. Dalalyan and A. Karagulyan. "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient". *Stochastic Processes and their Applications* (2019).

[DR17] G. K. Dziugaite and D. M. Roy. "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data". In: *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. 2017.

[DR18a] G. K. Dziugaite and D. M. Roy. "Data-dependent PAC-Bayes priors via differential privacy". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8430–8441.

[DR18b] G. K. Dziugaite and D. M. Roy. "Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of Entropy-SGD and data-dependent priors". In: *International Conference on Machine Learning*. 2018, pp. 1376–1385.

[Dur+17] A. Durmus, S. Le Corff, E. Moulines, and G. O. Roberts. "Optimal scaling of the random walk Metropolis algorithm under L^p mean differentiability". *Journal of Applied Probability* 54.4 (2017), pp. 1233–1260.

[Dwi+18] R. Dwivedi, Y. Chen, M. J. Wainwright, and B. Yu. "Log-concave sampling: Metropolis-Hastings algorithms are fast!" In: *Conference On Learning Theory*. 2018, pp. 793–797.

[Ebe14] A. Eberle. "Error bounds for Metropolis–Hastings algorithms applied to perturbations of Gaussian measures in high dimensions". *The Annals of Applied Probability* 24.1 (2014), pp. 337–377.

[Eft+16] C. Efthymiou, T. P. Hayes, D. Štefankovic, E. Vigoda, and Y. Yin. "Convergence of MCMC and loopy BP in the tree uniqueness region for the hard-core model". In: *Foundations of Computer Science (FOCS), 2016 IEEE 57th Annual Symposium on*. IEEE. 2016, pp. 704–713.

[EK86] S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and Convergence*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. Characterization and convergence. John Wiley & Sons, Inc., New York, 1986, pp. x+534. ISBN: 0-471-08186-8.

[Erv14] T. van Erven. *PAC-Bayes Mini-tutorial: A Continuous Union Bound*. 2014. arXiv: 1405.1580.

[Fer67] T. S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Probability and Mathematical Statistics, Vol. 1. Academic Press, New York-London, 1967, pp. xi+396.

[FHJ08] J. M. Flegal, M. Haran, and G. L. Jones. "Markov chain Monte Carlo: Can we trust the third significant figure?" *Statistical Science* (2008), pp. 250–260.

[Fus05] T. Fushiki. "Bootstrap prediction and Bayesian prediction under misspecified models". *Bernoulli* (2005), pp. 747–758.

[Ger+15] P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and J.-F. Roy. "Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm". *The Journal of Machine Learning Research* 16.1 (2015), pp. 787–860.

[Ger+16] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. "PAC-Bayesian theory meets Bayesian inference". In: *Advances in Neural Information Processing Systems*. 2016, pp. 1884–1892.

[Gho11] M. Ghosh. "Objective priors: An introduction for frequentists". *Statistical Science* (2011), pp. 187–202.

[GK06] E. Giné and V. Koltchinskii. "Concentration inequalities and asymptotic results for ratio type empirical processes". *The Annals of Probability* 34.3 (2006), pp. 1143–1216.

[GLX06] E. I. George, F. Liang, and X. Xu. "Improved minimax predictive densities under Kullback-Leibler loss". *The Annals of Statistics* (2006), pp. 78–91.

[GLX12] E. I. George, F. Liang, and X. Xu. "From minimax shrinkage estimation to minimax shrinkage prediction". *Statistical Science* (2012), pp. 82–94.

[GM19] P. D. Grünwald and N. A. Mehta. "A tight excess risk bound via a unified PAC-Bayesian–Rademacher–Shtarkov–MDL complexity". In: *Algorithmic Learning Theory*. 2019, pp. 433–465. arXiv: 1710.07732.

[GPW01] E. Gutierrez-Pena and S. G. Walker. "A Bayesian predictive approach to model selection". *Journal of Statistical Planning and Inference* 93.1 (2001), pp. 259–276.

[GR92] A. Gelman and D. B. Rubin. "Inference from iterative simulation using multiple sequences". *Statistical Science* (1992), pp. 457–472.

[GR98] C. Goutis and C. P. Robert. "Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections". *Biometrika* 85.1 (1998), pp. 29–37.

[Gra90] R. M. Gray. *Entropy and information*. Springer, 1990.

[GRS95] W. R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC press, 1995.

[Grü12] P. Grünwald. "The safe Bayesian". In: *International Conference on Algorithmic Learning Theory*. Springer. 2012, pp. 169–183.

[Gue19] B. Guedj. *A primer on PAC-Bayesian learning*. 2019. arXiv: 1901.05353.

[Han16] S. Hanneke. "Refined error bounds for several learning algorithms". *The Journal of Machine Learning Research* 17.1 (2016), pp. 4667–4721.

[Har98] J. Hartigan. "The maximum likelihood prior". *The annals of statistics* 26.6 (1998), pp. 2083–2103.

[Has70] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". *Biometrika* 57.1 (1970), pp. 97–109.

[HM02] P. Hajłasz and J. Malý. "Approximation in Sobolev spaces of nonlinear expressions involving the gradient". *Arkiv för Matematik* 40.2 (2002), pp. 245–274. ISSN: 1871-2487.

[HMS11] M. Hairer, J. C. Mattingly, and M. Scheutzow. "Asymptotic coupling and a general form of Harris' theorem with applications to stochastic delay equations". *Probability Theory and Related Fields* 149.1-2 (2011), pp. 223–259.

[Hub67] P. J. Huber. "The behavior of maximum likelihood estimates under nonstandard conditions". In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 1. 1967, pp. 221–233.

[HW13] P. D. Hoff and J. C. Wakefield. "Bayesian sandwich posteriors for pseudo-true parameters. A discussion of" Bayesian inference with misspecified models" by Stephen Walker." *Journal of Statistical Planning and Inference* 143.10 (2013), pp. 1638–1642.

[HY15] S. Hanneke and L. Yang. "Minimax analysis of active learning". *The Journal of Machine Learning Research* 16.1 (2015), pp. 3487–3602.

[Jer+04] M. Jerrum, J.-B. Son, P. Tetali, and E. Vigoda. "Elementary bounds on Poincaré and log-Sobolev constants for decomposable Markov chains". *Annals of Applied Probability* (2004), pp. 1741–1765.

[JH01] G. L. Jones and J. P. Hobert. "Honest exploration of intractable probability distributions via Markov chain Monte Carlo". *Statistical Science* (2001), pp. 312–334.

[JH04] G. L. Jones and J. P. Hobert. "Sufficient burn-in for Gibbs samplers for a hierarchical random effects model". *The Annals of Statistics* 32.2 (2004), pp. 784–817.

[JLM14] B. Jourdain, T. Lelièvre, and B. a. Miasojedow. "Optimal scaling for the transient phase of Metropolis Hastings algorithms: the longtime behavior". *Bernoulli* 20.4 (2014), pp. 1930–1978.

[JLM15] B. Jourdain, T. Lelièvre, and B. Miasojedow. "Optimal scaling for the transient phase of the random walk Metropolis algorithm: The mean-field limit". *The Annals of Applied Probability* 25.4 (2015), pp. 2263–2300.

[Jor04] M. I. Jordan. "Graphical Models". *Statistical Science* 19.1 (2004), pp. 140–155.

[KH13] K. Khare and J. P. Hobert. "Geometric ergodicity of the Bayesian lasso". *Electronic Journal of Statistics* 7 (2013), pp. 2150–2163.

[Kol06] V. Koltchinskii. "Local Rademacher complexities and oracle inequalities in risk minimization". *The Annals of Statistics* 34.6 (2006), pp. 2593–2656.

[Kom06] F. Komaki. "Shrinkage priors for Bayesian prediction". *the Annals of Statistics* (2006), pp. 808–819.

[Kom96] F. Komaki. "On asymptotic properties of predictive distributions". *Biometrika* (1996), pp. 299–313.

[KOS18] J. Kuntz, M. Ottobre, and A. M. Stuart. "Non-stationary phase of the MALA algorithm". *Stochastics and Partial Differential Equations: Analysis and Computations* 6.3 (2018), pp. 446–499.

[KOS19] J. Kuntz, M. Ottobre, and A. M. Stuart. "Diffusion limit for the random walk Metropolis algorithm out of stationarity". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 55.3 (Aug. 2019), pp. 1599–1648.

[KP02] V. Koltchinskii and D. Panchenko. "Empirical margin distributions and bounding the generalization error of combined classifiers". *The Annals of Statistics* 30.1 (2002), pp. 1–50.

[KST08] S. M. Kakade, K. Sridharan, and A. Tewari. "On the complexity of linear prediction: risk bounds, margin bounds, and regularization". In: *Advances in Neural Information Processing Systems*. 2008, pp. 793–800.

[Kv06] B. J. Kleijn and A. W. van der Vaart. "Misspecification in infinite-dimensional Bayesian statistics". *The Annals of Statistics* (2006), pp. 837–877.

[Kv12] B. Kleijn and A. van der Vaart. "The Bernstein-von-Mises theorem under misspecification". *Electronic Journal of Statistics* 6 (2012), pp. 354–381.

[Lac+07] A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. "PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier". In: *Advances in Neural Information Processing Systems*. 2007, pp. 769–776.

[Lan05] J. Langford. "Tutorial on practical prediction theory for classification". *Journal of Machine Learning Research* 6.Mar (2005), pp. 273–306.

[LeC53] L. LeCam. "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates". *Univ. California Publ. Statist.* 1 (1953), pp. 277–329.

[Lin99] J. Lindsey. "Some statistical heresies". *The Statistician* (1999), pp. 1–40.

[Liu08] J. S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science & Business Media, 2008.

[LLST13] G. Lever, F. Laviolette, and J. Shawe-Taylor. "Tighter PAC-Bayes bounds through distribution-dependent priors". *Theoretical Computer Science* 473 (2013), pp. 4–28.

[LM12] G. Lecué and C. Mitchell. "Oracle inequalities for cross-validation type procedures". *Electronic Journal of Statistics* 6 (2012), pp. 1803–1837.

[LM13] G. Lecué and S. Mendelson. *Learning subgaussian classes: Upper and minimax bounds*. 2013. arXiv: 1305.4825.

[Lon17] B. London. "A PAC-Bayesian analysis of randomized learning with application to stochastic gradient descent". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2931–2940.

[LRS15] T. Liang, A. Rakhlin, and K. Sridharan. "Learning with square loss: Localization through offset Rademacher complexity". In: *Conference on Learning Theory*. 2015, pp. 1260–1285.

[LV03] L. Lovász and S. Vempala. "Hit-and-run is fast and fun". *preprint, Microsoft Research* (2003).

[LV06] L. Lovász and S. Vempala. "Hit-and-run from a corner". *SIAM Journal on Computing* 35.4 (2006), pp. 985–1005.

[MARS19] F. Medina-Aguayo, D. Rudolf, and N. Schweizer. "Perturbation bounds for Monte Carlo within Metropolis via restricted approximations". *Stochastic Processes and their Applications* (2019).

[McA13] D. A. McAllester. *A PAC-Bayesian Tutorial with A Dropout Bound*. 2013. arXiv: 1307.2118.

[McA99] D. A. McAllester. "PAC-Bayesian Model Averaging". In: *Conference on Learning Theory*. 1999, pp. 164–170.

[MD15] J. W. Miller and D. B. Dunson. "Robust Bayesian inference via coarsening". *arXiv preprint arXiv:1506.06101* (2015).

[Men14] S. Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. 2014, pp. 25–39.

[Men17] S. Mendelson. ""Local" vs. "global" parameters–breaking the Gaussian complexity barrier". *The Annals of Statistics* 45.5 (2017), pp. 1835–1862.

[Met+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equation of state calculations by fast computing machines". *The Journal of Chemical Physics* 21.6 (1953), pp. 1087–1092.

[MN06] P. Massart and É. Nédélec. "Risk bounds for statistical learning". *The Annals of Statistics* 34.5 (2006), pp. 2326–2366.

[MPS12] J. C. Mattingly, N. S. Pillai, and A. M. Stuart. "Diffusion limits of the random walk Metropolis algorithm in high dimensions". *The Annals of Applied Probability* 22.3 (2012), pp. 881–930.

[MR00] R. A. Martin and D. Randall. "Sampling adsorbing staircase walks using a new Markov chain decomposition method". In: *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*. IEEE. 2000, pp. 492–502.

[MS17] O. Mangoubi and A. Smith. "Rapid Mixing of Hamiltonian Monte Carlo on Strongly Log-Concave Distributions". *arXiv preprint arXiv:1708.07114* (2017).

[MT12] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[MT94] S. P. Meyn and R. L. Tweedie. "Computable bounds for geometric convergence rates of Markov chains". *The Annals of Applied Probability* (1994), pp. 981–1011.

[MT96] K. L. Mengersen and R. L. Tweedie. "Rates of convergence of the Hastings and Metropolis algorithms". *The Annals of Statistics* 24.1 (1996), pp. 101–121.

[Mül13] U. K. Müller. "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix". *Econometrica* 81.5 (2013), pp. 1805–1849.

[NBS17] B. Neyshabur, S. Bhojanapalli, and N. Srebro. *A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks*. 2017. arXiv: 1707.09564.

[Ney+17] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. "Exploring generalization in deep learning". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5947–5956.

[NR06] P. Neal and G. Roberts. "Optimal scaling for partially updating MCMC algorithms". *The Annals of Applied Probability* 16.2 (2006), pp. 475–515.

[NR08] P. Neal and G. Roberts. "Optimal scaling for random walk Metropolis on spherically constrained target densities". *Methodology and Computing in Applied Probability* 10.2 (2008), pp. 277–297.

[NR11] P. Neal and G. Roberts. "Optimal scaling of random walk Metropolis algorithms with non-Gaussian proposals". *Methodology and Computing in Applied Probability* 13.3 (2011), pp. 583–601.

[NRY12] P. Neal, G. Roberts, and W. K. Yuen. "Optimal scaling of random walk Metropolis algorithms with discontinuous target densities". *The Annals of Applied Probability* 22.5 (2012), pp. 1880–1927.

[PG10] C. Pasarica and A. Gelman. "Adaptively Scaling the Metropolis Algorithm using Expected Squard Jumped Distance". *Statistica Sinica* 20.1 (2010), pp. 343–364.

[PST12] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions". *The Annals of Applied Probability* 22.6 (Dec. 2012), pp. 2320–2356.

[PV17] J. Piironen and A. Vehtari. "Comparison of Bayesian predictive methods for model selection". *Statistics and Computing* 27.3 (2017), pp. 711–735.

[QH17] Q. Qin and J. P. Hobert. "Asymptotically Stable Drift and Minorization for Markov Chains with Application to Albert and Chib's Algorithm". *arXiv preprint arXiv:1712.08867* (2017).

[RC04] C. P. Robert and G. Casella. "Monte Carlo Statistical Methods". *Springer, New York* (2004).

[RGG97] G. O. Roberts, A. Gelman, and W. R. Gilks. "Weak convergence and optimal scaling of random walk Metropolis algorithms". *The Annals of Applied Probability* 7.1 (1997), pp. 110–120.

[Ros02] J. S. Rosenthal. "Quantitative convergence rates of Markov chains: A simple account". *Electronic Communications in Probability* 7 (2002), pp. 123–128.

[Ros11] J. S. Rosenthal. "Optimal proposal distributions and adaptive MCMC". *Handbook of Markov Chain Monte Carlo* 4 (2011).

[Ros95a] J. S. Rosenthal. "Minorization conditions and convergence rates for Markov chain Monte Carlo". *Journal of the American Statistical Association* 90.430 (1995), pp. 558–566.

[Ros95b] J. S. Rosenthal. "Rates of convergence for Gibbs sampling for variance component models". *The Annals of Statistics* (1995), pp. 740–761.

[Ros96] J. S. Rosenthal. "Analysis of the Gibbs sampler for a model related to James-Stein estimators". *Statistics and Computing* 6.3 (1996), pp. 269–275.

[RR01] G. O. Roberts and J. S. Rosenthal. "Optimal scaling for various Metropolis–Hastings algorithms". *Statistical science* 16.4 (2001), pp. 351–367.

[RR14] G. O. Roberts and J. S. Rosenthal. "Minimising MCMC variance via diffusion limits, with an application to simulated tempering". *The Annals of Applied Probability* 24.1 (2014), pp. 131–149.

[RR16] G. O. Roberts and J. S. Rosenthal. "Complexity bounds for Markov chain Monte Carlo algorithms via diffusion limits". *Journal of Applied Probability* 53.2 (2016), pp. 410–420.

[RR98] G. O. Roberts and J. S. Rosenthal. "Optimal scaling of discrete approximations to Langevin diffusions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60.1 (1998), pp. 255–268.

[RS15] B. Rajaratnam and D. Sparks. "MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains". *arXiv preprint arXiv:1508.00947* (2015).

[RS18] D. Rudolf and B. Sprungk. "On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm". *Foundations of Computational Mathematics* 18.2 (2018), pp. 309–343.

[RS94] G. O. Roberts and A. F. Smith. "Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms". *Stochastic processes and their applications* 49.2 (1994), pp. 207–216.

[RT03] R. Royall and T.-S. Tsou. "Interpreting statistical evidence by using imperfect models: robust adjusted likelihood functions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 391–404.

[RT99] G. O. Roberts and R. L. Tweedie. "Bounds on regeneration times and convergence rates for Markov chains". *Stochastic Processes and their applications* 80.2 (1999), pp. 211–229.

[Sch96] M. Schervish. *Theory of Statistics*. Springer Series in Statistics. Springer New York, 1996. ISBN: 9780387945460.

[SD12] N. Shephard and A. Doucet. *Robust inference on parameters via particle filters and sandwich covariance matrices*. Tech. rep. 2012.

[SDG06] T. J. Sweeting, G. S. Datta, and M. Ghosh. "Nonsubjective priors via predictive relative entropy regret". *The Annals of Statistics* (2006), pp. 441–468.

[SFR10] C. Sherlock, P. Fearnhead, and G. O. Roberts. "The random walk Metropolis: linking theory and practice through a case study". *Statistical Science* (2010), pp. 172–190.

[She+15] C. Sherlock, A. H. Thiery, G. O. Roberts, and J. S. Rosenthal. "On the efficiency of pseudo-marginal random walk Metropolis algorithms". *The Annals of Statistics* 43.1 (2015), pp. 238–275.

[She06] C. Sherlock. "Methodology for inference on the Markov modulated Poisson process and theory for optimal scaling of the random walk Metropolis". PhD thesis. Lancaster University, 2006.

[SJ89] A. Sinclair and M. Jerrum. "Approximate counting, uniform generation and rapidly mixing Markov chains". *Information and Computation* 82.1 (1989), pp. 93–133.

[SL17] S. L. Smith and Q. V. Le. *A Bayesian perspective on generalization and stochastic gradient descent*. 2017. arXiv: 1710.06451.

[SR09] C. Sherlock and G. Roberts. "Optimal scaling of the random walk Metropolis on elliptically symmetric unimodal targets". *Bernoulli* (2009), pp. 774–798.

[Sta96] J. E. Stafford. "A robust adjustment of the profile likelihood". *The Annals of Statistics* 24.1 (1996), pp. 336–352.

[Ste81] C. M. Stein. "Estimation of the mean of a multivariate normal distribution". *The Annals of Statistics* 9.6 (1981), pp. 1135–1151. ISSN: 0090-5364.

[Stu10] A. M. Stuart. "Inverse problems: a Bayesian perspective". *Acta Numerica* 19 (2010), pp. 451–559.

[STW97] J. Shawe-Taylor and R. C. Williamson. "A PAC Analysis of a Bayesian Estimator". In: *Conference on Learning Theory*. 1997, pp. 2–9.

[Swe11] T. Sweeting. "Discussion of "Objective priors: an introduction for frequentists" by M. Ghosh [MR2858380]". *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 26.2 (2011), pp. 206–209. ISSN: 0883-4237.

[Taw17] N. Tawn. "Towards Optimality of the Parallel Tempering Algorithm". PhD thesis. University of Warwick, 2017.

[Thi+16] N. Thiemann, C. Igel, O. Wintenberger, and Y. Seldin. *A strongly quasiconvex PAC-Bayesian bound*. 2016. arXiv: 1608.05610.

[TS13] I. O. Tolstikhin and Y. Seldin. "PAC-Bayes-empirical-Bernstein inequality". In: *Advances in Neural Information Processing Systems*. 2013, pp. 109–117.

[Vem05] S Vempala. "Geometric random walk: a survey". *Combinatorial and computational geometry* 52 (2005), pp. 577–616.

[VO12] A. Vehtari and J. Ojanen. "A survey of Bayesian predictive methods for model assessment, selection and comparison". *Statistics Surveys* 6 (2012), pp. 142–228.

[Wai19] M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[Wal+99] S. G. Walker, P. Damien, P. W. Laud, and A. F. Smith. "Bayesian nonparametric inference for random distributions and related functions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3 (1999), pp. 485–527.

[Wal02] S. G. Walker. "Bayesian Nonparametric and Parametric Inference". *Journal of Iranian Statistical Society* 1.1–2 (2002), pp. 143–163.

[Wal49] A. Wald. "Note on the consistency of the maximum likelihood estimate". *The Annals of Mathematical Statistics* 20.4 (1949), pp. 595–601.

[Weg03] M. Wegkamp. "Model selection in nonparametric regression". *The Annals of Statistics* 31.1 (2003), pp. 252–273.

[Whi82] H. White. "Maximum likelihood estimation of misspecified models". *Econometrica: Journal of the Econometric Society* (1982), pp. 1–25.

[WSH09a] D. Woodard, S. Schmidler, and M. Huber. "Sufficient conditions for torpid mixing of parallel and simulated tempering". *Electronic Journal of Probability* 14 (2009), pp. 780–804.

[WSH09b] D. B. Woodard, S. C. Schmidler, and M. Huber. "Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions". *The Annals of Applied Probability* (2009), pp. 617–640.

[YR17] J. Yang and J. S. Rosenthal. "Complexity Results for MCMC derived from Quantitative Bounds". *arXiv:1708.00829* (2017).

[YRR19] J. Yang, G. O. Roberts, and J. S. Rosenthal. "Optimal scaling of Metropolis algorithms on general target distributions". *arXiv preprint arXiv:1904.12157* (2019).

[YSR19] J. Yang, S. Sun, and D. M. Roy. "Fast-rate PAC-Bayes Generalization Bounds via Shifted Rademacher Processes". In: *Advances in Neural Information Processing Systems*. 2019, pp. 10802–10812.

[ZBK17] G. Zanella, M. Bédard, and W. S. Kendall. "A Dirichlet form approach to MCMC optimal scaling". *Stochastic Processes and their Applications* 127.12 (2017), pp. 4053 –4082.

[ZC04] V. A. Zorich and R. Cooke. *Mathematical analysis II*. Springer Science & Business Media, 2004.

[ZH18] N. Zhivotovskiy and S. Hanneke. "Localization of VC classes: Beyond local Rademacher complexities". *Theoretical Computer Science* 742 (2018), pp. 27–49.

[van+15] T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. "Fast Rates in Statistical and Online Learning". *Journal of Machine Learning Research* 16 (2015), pp. 1793–1861.