Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo

Rong Ge[*], Holden Lee[†], Andrej Risteski[‡]

Abstract

A key task in Bayesian statistics is sampling from distributions that are only specified up to a partition function (i.e., constant of proportionality). However, without any assumptions, sampling (even approximately) can be #P-hard, and few works have provided "beyond worst-case" guarantees for such settings.

For log-concave distributions, classical results going back to Bakry and Émery [1985] show that natural continuous-time Markov chains called *Langevin diffusions* mix in polynomial time. The most salient feature of log-concavity violated in practice is unimodality: commonly, the distributions we wish to sample from are multimodal. In the presence of multiple deep and well-separated modes, Langevin diffusion suffers from torpid mixing.

We address this problem by combining Langevin diffusion with *simulated tempering*. The result is a Markov chain that mixes more rapidly by transitioning between different temperatures of the distribution. We analyze this Markov chain for the canonical multi-modal distribution: a mixture of Gaussians (of equal variance). The algorithm based on our Markov chain provably samples from distributions that are close to mixtures of gaussians, given access to the gradient of the log-pdf. For the analysis, we use a spectral decomposition theorem for graphs Gharan and Trevisan [2014] and a Markov chain decomposition technique Madras and Randall [2002].

The full version of the paper is available at `http://www.arxiv.org/abs/1710.02736` Ge et al. [2017].

1 Introduction

In recent years, one of the most fruitful directions of research has been providing theoretical guarantees for optimization in non-convex settings. In particular, a routine task in both unsupervised and supervised learning is fitting optimal parameters for a model in some parametric family. Theoretical successes in this context range from analyzing tensor-based approaches using method-of-moments, to iterative techniques like gradient descent, EM, and variational inference in a variety of models. These models include topic models Anandkumar et al. [2012], Arora et al. [2012, 2013], Awasthi and Risteski [2015], dictionary learning Arora et al. [2015], Agarwal et al. [2014], gaussian mixture models Hsu and Kakade [2013], and Bayesian networks Arora et al. [2017].

Max-likelihood values of unobserved quantities via optimization is reasonable in many learning settings, as when the number of samples is large max-likelihood will converge to the true value. However, for Bayesian inference problems (e.g. given a document, what topics is it about) the

[*]Duke University, Computer Science Department `rongge@cs.duke.edu`

[†]Princeton University, Mathematics Department `holdenl@princeton.edu`

[‡]Massachusetts Institute of Technology, Applied Mathematics and IDSS `risteski@mit.edu`

number of samples can be limited and maximum likelihood may not be well-behaved Sontag and Roy [2011]. In these cases we would prefer to *sample* from the posterior distribution. More generally, the scenario is sampling from the *posterior* distribution over the latent variables h of a latent variable Bayesian model $p(h, x) = p(h)p(x|h)$ whose parameters are known: the *posterior* distribution $p(h|x)$ has the form $p(h|x) = \frac{p(h)p(x|h)}{p(x)}$ which is easy to evaluate up to a constant of proportionality ($p(x)$), but without structural assumptions such distributions are often hard to sample from (exactly or approximately) even for simple models like topic models Sontag and Roy [2011].

The sampling analogues of convex functions (arguably the widest class of real-valued functions for which optimization is easy) are *log-concave* distributions, i.e. distributions of the form $p(x) \propto e^{-f(x)}$ for a convex function $f(x)$. Recently, there has been renewed interest in analyzing a popular Markov Chain for sampling from such distributions, when given gradient access to f—a natural setup for the posterior sampling task described above. A practically popular Markov chain called *Langevin Monte Carlo* has been proven to work, with various rates depending on the properties of f Dalalyan [2016], Durmus and Moulines [2016], Dalalyan [2017].

Log-concave distributions are necessarily uni-modal: their density functions have only one local maximum, which must then be a global maximum. This fails to capture many interesting scenarios. Even simple posterior distributions are neither log-concave nor uni-modal, for instance, the posterior distribution of the means for a mixture of Gaussians. Certainly complicated posterior distributions like the ones associated with deep generative modelsRezende et al. [2014] and variational auto-encoders Kingma and Welling [2013] are believed to be multimodal as well.

The goal of this work is to initiate an exploration of provable methods for sampling "beyond log-concavity," in parallel to optimization "beyond convexity". As worst-case results are prohibited by hardness results, we must again make assumptions on the distributions we will be interested in. As a first step, in this paper we consider the prototypical multimodal distribution, a mixture of Gaussians.

1.1 Our results

We formalize the problem of interest as follows. We wish to sample from a distribution $p : \mathbb{R}^d \to \mathbb{R}$, such that $p(x) \propto e^{-f(x)}$, and we are allowed to query $\nabla f(x)$ and $f(x)$ at any point $x \in \mathbb{R}^d$.

To start with, we focus on a problem where $e^{-f(x)}$ is the density function of a mixture of gaussians. That is, given centers $\mu_1, \mu_2, \ldots, \mu_n \in \mathbb{R}^d$, weights $w_1, w_2, \ldots, w_n$ ($\sum_{i=1}^n w_i = 1$), variance σ^2 (all the gaussians are spherical with same covariance matrix $\sigma^2 I$), the function $f(x)$ is defined as[1]

$$f(x) = -\log\left(\sum_{i=1}^{n} w_i \exp\left(-\frac{\|x-\mu_i\|^2}{2\sigma^2}\right)\right). \tag{1}$$

Furthermore, suppose that D is such that $\|\mu_i\| \leq D, \forall i \in [n]$. We show that there is an efficient algorithm that can sample from this distribution given just access to $f(x)$ and $\nabla f(x)$.

Theorem 1.1 (main, informal)**.** *Given $f(x)$ as defined in Equation (1), there is an algorithm with running time* $\mathrm{poly}(w_{\min}, D, d, 1/\varepsilon, 1/\sigma^2)$ *that outputs a sample from a distribution within TV-distance ε of $p(x)$.*

[1]Note that the expression inside the log is essentially the probability density of a mixture of gaussians, except the normalization factor is missing. However, the normalization factor can just introduce a constant shift of f and does not really change ∇f.

Note that the algorithm does *not* have direct access to $\mu_1, \mu_2, \ldots, \mu_n$, which makes sampling from this mixture of Gaussians distribution very non-trivial.

Of course, requiring the distribution to be *exactly* a mixture of gaussians is a very strong assumption. Our results can be generalized to all functions that are "close" to a mixture of Gaussians. More precisely, the function f satisfies the following properties:

$$\exists \tilde{f}: \quad \mathbb{R}^d \to \mathbb{R} \text{ where } \left\|\tilde{f} - f\right\|_\infty \leq \Delta \ , \ \left\|\nabla \tilde{f} - \nabla f\right\|_\infty \leq \tau \text{ and } \nabla^2 \tilde{f}(x) \preceq \nabla^2 f(x) + \tau I, \forall x \in \mathbb{R}^d \tag{2}$$

$$\text{and } \tilde{f}(x) = -\log\left(\sum_{i=1}^{n} w_i \exp\left(-\frac{\|x-\mu_i\|^2}{2\sigma^2}\right)\right) \tag{3}$$

Intuitively, these conditions show that the density of the distribution is within a e^{Δ} multiplicative factor to an (unknown) mixture of gaussians. Our theorem can be generalized to this case.

Theorem 1.2 (general case, informal)**.** *For function $f(x)$ that satisfies Equations (2) and (3), there is an algorithm that runs in time* $\text{poly}\left(w_{\min}, D, d, \frac{1}{\varepsilon}, e^{\Delta}, \tau\right)$ *that outputs a sample x from a distribution that has TV-distance at most ε from $p(x)$.*

1.2 Prior work

Our algorithm will use two classical techniques in the theory of Markov chains: *Langevin diffusion*, a chain for sampling from distributions in the form $p(x) \propto e^{-f(x)}$ given only gradient access to f and *simulated tempering*, a heuristic technique used for tackling multimodal distributions. We recall briefly what is known for both of these techniques.

For Langevin dynamics, convergence to the stationary distribution is a classic result Bhattacharya [1978]. Understanding the mixing time of the continuous dynamics for log-concave distributions is also classic: Bakry and Émery [1985], Bakry et al. [2008]. Of course, algorithmically, one can only run a "discretized" version of the Langevin dynamics and one wants to make sure this doesn't change the converged-to distribution and mixing time too much. Such results are more recent: Dalalyan [2016], Durmus and Moulines [2016], Dalalyan [2017] in the case of log-concave distributions and Raginsky et al. [2017] for arbitrary non-log-concave distributions with certain regularity and decay properties. Of course, the mixing time is exponential in general when the spectral gap of the chain is small; furthermore, it has long been known that transitioning between different modes can take an exponentially long time, a phenomenon known as meta-stability Bovier et al. [2002, 2004, 2005].

Clearly for distributions that are far from being log-concave and have many deep modes, additional techniques will be necessary. Among many proposed heuristics for such situations is simulated tempering, which effectively runs multiple Markov chains, each corresponding to a different temperature of the original chain, and "mixes" between these different Markov chains. The intuition is that the Markov chains at higher temperature can move between modes more easily, and if one can "mix in" points from these into the lower temperature chains, their mixing time ought to improve as well. Provable results of this heuristic are however few and far between. Woodard et al. [2009], Zheng [2003] lower-bound the spectral gap for generic simulated tempering chains. The crucial technique our paper shares with theirs is a Markov chain decomposition technique due to Madras

and Randall [2002]. However, for the scenario of Section 1.1 we are interested in, the spectral gap bound in Woodard et al. [2009] is exponentially small as a function of the number of modes. Our result will remedy this.

2 Our Algorithm

Our algorithm runs simulated tempering chain, with a polynomial number of temperatures, while running discretized Langevin dynamics at the various temperatures.

Given some Markov chain with a corresponding stationary distribution, *simulated tempering* constructs a new Markov chain whose state space is a product of the original state space and a temperature, and it runs "in parallel" the Markov chains corresponding to the stationary distribution at the different temperatures, while allowing "mixing" of points from different temperatures. Concretely, if $M_i, i \in [L]$ are the Markov chains corresponding to temperature i, we evolve a point $(x, k), k \in [L]$ as follows:

1. With probability $\frac{1}{2}$, keep k fixed, and update x according to M_k.
2. With probability $\frac{1}{2}$, do the following Metropolis-Hastings step: draw k' randomly from $\{0, \ldots, L-1\}$. Then transition to (x, k') with probability $\min\left\{\frac{p_{k'}(x)}{p_k(x)}, 1\right\}$.

3 Overview of proof

We will sketch the proof for the purposes of this extended abstract. The key parts of the proof concern two *decomposition theorems* for bounding the mixing time of Markov Chains:

- *Decomposition theorem for simulated tempering*: we prove the following statement—if we can partition the state space corresponding to each temperature, such that (1) the Markov chain at that temperature mixes fast in each of the blocks and (2) each block doesn't have too small of a measure under the stationary measure for that temperature, then we can bound the mixing time of the simulated tempering chain.
- *Decomposing the state space for mixtures of Gaussians*: towards applying the above result, we produce such a decomposition for mixtures of Gaussians (when the individual chains in the simulated tempering are Langevin dynamics). Concretely, we show that for mixtures of n Gaussians, Langevin dynamics exhibits a gap before the $(n+1)$-st eigenvalue. From this, we use spectral techniques by Gharan and Trevisan [2014] to partition the state space in each temperature into at most n blocks, where each block is well-connected inside, and blocks are poorly "interconnected." Finally, for mixtures of Gaussians, we prove small sets have a good expansion under Langevin dynamics, so cannot be blocks in the decomposition provided by the algorithm by Gharan and Trevisan [2014]. This provides the necessary decomposition.

Finally, there are a few technical details to work out. These include adapting discrete time/space results such as Gharan and Trevisan [2014] to the continuous time/space setting of Langevin dynamics, proving the Markov chain mixes at the highest temperature, proving the discretized Markov chain approximates the continuous time Markov chain, proving the partition functions are estimated correctly which allows us to run the simulated tempering chain, and proving the L^∞ perturbation tolerance.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning sparsely used overcomplete dictionaries. In *Conference on Learning Theory (COLT)*, 2014.

Anima Anandkumar, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond svd. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2012.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pages 280–288, 2013.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. 2015.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski. Provable learning of noisy-or networks. In *Symposium on the Theory of Computing (STOC)*, 2017.

Pranjal Awasthi and Andrej Risteski. On some provably correct cases of variational inference for topic models. In *Advances in Neural Information Processing Systems*, pages 2098–2106, 2015.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*, pages 177–206. Springer, 1985.

Dominique Bakry, Franck Barthe, Patrick Cattiaux, and Arnaud Guillin. A simple proof of the poincaré inequality for a large class of probability measures including the log-concave case. *Electron. Commun. Probab*, 13:60–66, 2008.

RN Bhattacharya. Criteria for recurrence and existence of invariant measures for multidimensional diffusions. *The Annals of Probability*, pages 541–553, 1978.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability and low lying spectra in reversible markov chains. *Communications in mathematical physics*, 228(2):219–255, 2002.

Anton Bovier, Michael Eckhoff, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes i: Sharp asymptotics for capacities and exit times. *Journal of the European Mathematical Society*, 6(4):399–424, 2004.

Anton Bovier, Véronique Gayrard, and Markus Klein. Metastability in reversible diffusion processes ii: Precise asymptotics for small eigenvalues. *Journal of the European Mathematical Society*, 7 (1):69–99, 2005.

Arnak S Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Arnak S Dalalyan. Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent. *arXiv preprint arXiv:1704.04752*, 2017.

Alain Durmus and Eric Moulines. High-dimensional bayesian inference via the unadjusted langevin algorithm. 2016.

Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. *arXiv preprint arXiv:1710.02736*, 2017.

Shayan Oveis Gharan and Luca Trevisan. Partitioning into expanders. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1256–1266. Society for Industrial and Applied Mathematics, 2014.

Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Neal Madras and Dana Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, pages 581–606, 2002.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *arXiv preprint arXiv:1702.03849*, 2017.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

David Sontag and Dan Roy. Complexity of inference in latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1008–1016, 2011.

Dawn B Woodard, Scott C Schmidler, and Mark Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *The Annals of Applied Probability*, pages 617–640, 2009.

Zhongrong Zheng. On swapping and simulated tempering algorithms. *Stochastic Processes and their Applications*, 104(1):131–154, 2003.