Fast Convergence for Langevin Diffusion with Matrix Manifold Structure

February 18, 2020

Abstract

In this paper, we study the problem of sampling from distributions of the form $p(x) = e^{-\beta f(x)}/Z$ where Z is the normalizing constant and β is the inverse temperature, for some function f whose values and gradients we can query. This mode of access to f is natural in the scenarios in which such problems arise, for instance sampling from posteriors in parametric Bayesian models and optimizing certain PAC-Bayes bounds on the generalization error. Classical results (Bakry and Émery, 1985) show that a natural random walk, the Langevin dynamics, mixes rapidly when f is convex. Unfortunately, even in simple examples, the applications listed above will entail working with functions f that are nonconvex — for which sampling from p may in general require an exponential number of queries (Ge et al., 2018).

In this paper, we study one aspect of nonconvexity relevant for modern machine learning applications: existence of invariances (symmetries) in the function f, as a result of which the distribution p will have *manifolds* of points with equal probability. We give a recipe for proving mixing time bounds of Langevin dynamics in order to sample from manifolds of *local optima* of the function f in settings where the distribution is well-concentrated around them. We specialize our arguments to classic matrix factorization-like Bayesian inference problems where we get noisy measurements $\mathcal{A}(XX^T), X \in \mathbb{R}^{d\times k}$ of a low-rank matrix, for a linear "measurements" operator $\mathcal{A}$—thus $f(X) = \|\mathcal{A}(XX^T) - b\|_2^2, X \in \mathbb{R}^{d\times k}$, and β the inverse of the standard deviation of the noise. Such functions f are invariant under orthogonal transformations, and include problems like matrix factorization ($\mathcal{A}$ is the identity map), matrix sensing ($\mathcal{A}$ collects the measurements), matrix completion ($\mathcal{A}$ is the projection operator to the visible entries). Beyond sampling, Langevin dynamics is a popular toy model for studying stochastic gradient descent. Along these lines, we believe that our work is an important first step towards understanding how SGD behaves when there is a high degree of symmetry in the space of parameters the produce the same output.

The full paper can be found on ArXiv.

1 Introduction

1.1 Background

In this paper, we study the problem of sampling from a distribution $p(X) = \frac{e^{-\beta f(X)}}{Z}$ where Z is the normalizing constant, for some particular families of functions $f(X)$ that are *nonconvex*, and we can access f through a value and gradient oracle. This problem is the sampling equivalent to the classical setup of minimizing a function f, given access to the same oracles, which is the usual sandbox in which query complexity of optimization can be quantified precisely.

Mirroring what happens for optimization, when $f(X)$ is convex (i.e. $p(X)$ is logconcave), there are a variety of algorithms for efficiently sampling from $p(X)$. Beyond that, however, the problem is in general hard: Ge et al. (2018) prove an exponential lower bound on the number of queries required. Nevertheless, the non-logconcave case is relevant in practice because of its wide-ranging applications:

1. **Bayesian inference:** In instances when we have a prior on a random variable X, of which we get noisy observations, the posterior distribution often takes the above form and is called a Gibbs distribution. Also

β is called the inverse temperature and depends on the level of noise in the model. Intuitively when β is large, the distribution places more weight on the X's that are close to the observation as measured by $f(X)$. And when β is small, it samples from a larger entropy distribution around the observation.

We will consider several natural instances in this paper, where we get "measurements" $\mathcal{A}(XX^T)$ of a low-rank matrix, perturbed by some amount of noise—subsuming problems like noisy matrix factorization, matrix sensing, matrix completion.

A more complicated version of this are latent variable models, in which an observable random variable Y has an explicit form, conditioned on some latent variable X. The inference task of sampling the posterior distribution of the latent variables X, namely $P(X|Y)$, by Bayes law, will be captured by our setup, as $P(X|Y) \propto P(X)P(Y|X)$. These posteriors are not log-concave, even for models as simple as mixtures of Gaussians.

2. **Exploring complicated loss surfaces:** Many modern machine learning models (e.g. deep neural networks) have a high degree of symmetry in the space of their parameters that produce the same output. When $f(X)$ measures the error of the parameters on the set of training examples, the set of local minima of the loss is often an implicitly defined manifold because of these symmetries. Being able to sample from these manifolds can be beneficial in the context of interpretability, allowing us to discover which samples the loss treats similarly or to otherwise explore the geometry of the loss function (Zeiler and Fergus, 2014). For nearly all instances of interest, $f(X)$ is highly non-convex, so the distribution we are sampling from will not be log-concave.
3. **Improving generalization:** Recent works attempting to understand generalization in deep learning suggest that, when $f(X)$ is the loss function of an overparameterized model, sampling from a set with larger volume or entropy has benefits in terms of the generalization guarantees. In particular, it optimizes a PAC Bayes generalization bound (Dziugaite and Roy, 2017b,a). Further experimental evidence supporting this hypothesis was given by Shwartz-Ziv and Tishby (2017) who showed that in synthetic experiments a large fraction of the time training a deep neural network is spent diffusing along a plateau of the loss – i.e. stochastic gradient descent does not decrease the value of the loss, but spreads out on a level set. They argue that this increases the entropy of the produced output and improves generalization.

   Similarly as above, the resulting distributions we wish to sample from will not be log-concave.

In the hopes of exploring the landscape of tractable distributions we can sample from, for which $f(X)$ is nonconvex, we ask:

Question. *Are there simple and (statistically) meaningful families of nonconvex functions $f(X)$ where we can provably sample from $p(X)$ in polynomial time?*

The aspect of $f(X)$ we wish to capture in this paper is the existence of *symmetries*, motivated by applications (2) and (3) above. Taking inspiration from the literature on nonconvex optimization, we consider the case when f is the objective corresponding to relatives of *noisy low rank matrix factorization*, which is invariant under *orthogonal transforms*—e.g. *matrix completion*, *matrix sensing*, *robust PCA*.

When we can query the values and gradients of $f(X)$, there is a natural algorithm for sampling from $p(X)$ called Langevin dynamics. In its continuous form, it is described by the following stochastic differential equation:

$$dX_t = -\beta \nabla f(X_t)dt + dB_t$$

where B_t is Brownian motion of the appropriate dimension. It is well known that under mild conditions on $f(X)$, the stationary distribution is indeed $p(X)$ (Bhattacharya, 1978). When $p(X)$ is log-concave it is known that the Langevin dynamics mixes quickly (Bakry and Émery, 1985; Bakry et al., 2008; Bubeck et al., 2015).

1.2 Our Results

In this work, we study the problem of sampling from $p(X)$ when

$$f(X) = \|\mathcal{A}(XX^T) - b\|_2^2 \tag{1}$$

where X is a $d \times k$ matrix, $\mathcal{A}$ is a linear measurements operator, s.t.

$$\forall i \in [L], M \in \mathbb{R}^{d\times d}, \mathcal{A}(M)_i = \mathrm{Tr}(A_i^T M), A_i \in \mathbb{R}^{d\times d} \tag{2}$$

and b_i are noisy measurements of some ground-truth matrix, namely

$$\forall i \in [L], b_i = \mathrm{Tr}(A_i^T M^*) + n_i \tag{3}$$

where $M^* = X^*(X^*)^T \in \mathbb{R}^{d\times d}$ is of rank k with $\sigma_{\max}, \sigma_{\min}$ denoting the largest and smallest singular values of X^* respectively, and let $\kappa = \frac{\sigma_{\max}}{\sigma_{\min}}$ denote the condition number. Furthermore, $n_i \sim N(0, \frac{1}{\beta})$—i.e. Gaussian noise with variance $\frac{1}{\beta}$.

In this noise model, $p(X) \propto e^{-\beta f(X)}$ is exactly the posterior distribution over X.

We will consider three instances of $\mathcal{A}$:

1. **Noisy matrix factorization**: $\mathcal{A}$ is simply the identity operator, i.e. $\mathcal{A}(XX^T) = \mathrm{vec}(XX^T)$.
2. **Matrix sensing** with measurements satisfying *restricted isometry (RIP)*: $\mathcal{A}$ satisfies

   $$\left(1 - \frac{1}{20}\right)\|X\|_F^2 \le \|\mathcal{A}(X)\|_2^2 \le \left(1 + \frac{1}{20}\right)\|X\|_F^2$$

   holds for all matrices $X \in \mathbb{R}^{d\times d}$ of rank at most $2k$.
3. **Matrix completion**: $\mathcal{A}$ is a projection to a set of randomly chosen entries $\Omega \subseteq [d] \times [d]$, namely $\mathcal{A} = P_\Omega$, where $P_\Omega(Z)_{i,j} = P_{i,j}Z_{i,j}$, with $P_{i,j} = 1$ if $(i,j) \in \Omega$ and 0 otherwise. Furthermore, the probability of sampling an entry is $p = \Omega\left(\max(\mu^6\kappa^{16}k^4, \mu^4\kappa^4k^6)\frac{\log^2 d}{d}\right)$, where μ is an upper bound on the incoherence of M^*, that is the singular value decomposition $M^* = U\Sigma V^T$ satisfies $\max_{i\in[d]} \|e_i^T U\| \le \sqrt{\mu\frac{k}{d}}$.

We note that in each corresponding context, the assumptions are the standard ones in the literature on non-convex optimization – so our results can be viewed as sampling analogues of these results. We furthermore note that we chose the Gaussian noise setting in order to have a sampling problem from a natural posterior – however, from an algorithmic point of view, even the setting where $b = \mathcal{A}(X^*(X^*)^T)$, and we wish to sample from the corresponding p is equally hard/interesting, as the distribution is not log-concave, and satisfies the same manifold structure.

We will prove that Langevin dynamics mixes in polynomial time when β is at least a fixed polynomial in d, k and the condition number of M. Our analysis is geometric in nature and is based on a suitable decomposition of $p(X)$ along something akin to the level sets of $f(X)$. We prove various differential geometric estimates on the curvatures and distribution of volume along these level sets, and combine these to prove a restricted Poincaré inequality. In fact, our strategy is quite generic and can be reformulated as a general recipe that might be possible to follow in even more technically challenging settings.

Our first contribution is to formalize this general recipe. We study the general problem of sampling from the conditional distribution close to a manifold $\mathbf{M}$ that is a level set of $f(X)$ and has the property that all of its points are local minima – i.e. for all $X \in \mathbf{M}$ we have $\nabla f(X) = 0, \nabla^2 f(X) \succeq 0$ and $f(X) = s_0$.

Towards stating the result informally, consider an arbitrary point $X_0 \in \mathbf{M}$, and denote the "norm-bounded" normal space at X_0:

$$\mathbf{B} = \{\Delta : \Delta \in N_{X_0}(\mathbf{M}), \|\Delta\|_2 \le s\}$$

Furthermore, we assume that $\forall X \in \mathbf{M}$, there is a differential bijection

$$\phi_X : \mathbf{B} \to \{\Delta : \Delta \in N_X(\mathbf{M}), \|\Delta\|_2 \le s\}$$

that "transports" the normal space at X_0 to the normal space at X. With this in mind, it's natural to consider the "level set" corresponding to Δ, namely

$$\mathbf{M}^\Delta := \{X + \phi_X(\Delta) : X \in \mathbf{M}\}$$

Finally let $\tilde{p}^\Delta(X)$ denote the restriction of $p(X)$ to $\mathbf{M}^\Delta$ (with a suitable change of measure correction that comes from the coarea formula) and let $q(\Delta)$ denote the total weight that $p(X)$ places on each $\mathbf{M}^\Delta$ (with the same correction, again coming from the coarea formula). We show the following:

Theorem 1 (Informal)**.** *Suppose the following conditions hold:*

(1) (Nearness to the manifold): When initialized close to $\mathbf{M}$*, the Langevin dynamics stay within distance* η *from* $\mathbf{M}$ *up to time* T *with high probability.*

(2) (Poincaré inequality along level sets): The distributions $\tilde{p}^{\Delta}$ *for all* $\Delta \in \mathbf{B}$ *have a Poincaré constant bounded by* C_{level}

(3) (Poincaré inequality across level sets): The distribution q *has a Poincaré constant bounded by* C_{across}*.*

(4) (Bounded change of manifold probability): If we denote by $G_{\Delta} : \mathbf{M} \to \mathbf{M}^{\Delta}$ *the map* $G_{\Delta}(X) = X + \phi_X(\Delta)$*, for all* $X \in \mathbf{M}$ *and* $\Delta \in \mathbf{B}$*, the relative change (with respect to* Δ*) in the manifold density is bounded*[1]*:*

$$\left\| \frac{\nabla_{\mathbf{B}} \left(p^{\Delta}(X + \phi_X(\Delta)) det\left((dG_{\Delta})_X\right)\right)}{p^{\Delta}(X + \phi_X(\Delta)) det\left((dG_{\Delta})_X\right)} \right\|_2 \leq C_{change}$$

Then Langevin dynamics run for time

$$O\left(\max\left(1, C_{level}\right) \max\left(1, C_{across}\right) \max\left(1, C_{change}^2\right)\right)$$

outputs a sample from a distribution that is close in total variation distance to the conditional distribution of $p(X)$ *restricted to* $\mathcal{D}$ *with high probability.*

Our second main contribution is in showing that each of these conditions can be proven to hold in the setting of matrix factorization-like functions f. Namely, when $f(X) = \|\mathcal{A}(XX^T) - b\|_F^2$, for popular choices of operators $\mathcal{A}$ (giving rise to matrix factorization, matrix sensing and matrix completion), the set of global minimizers take the form

$$\mathbf{E}_1 = \{X^*R, R \in O(k), \det(R) = 1\} \text{ and } \mathbf{E}_2 = \{X^*R, R \in O(k), \det(R) = -1\}$$

where X^* is any fixed minimum of $f(X)$. In general, it will take exponentially long for Langevin dynamics to transition from one manifold to the other. However we show that it successfully discovers one of them and samples from $p(X)$ restricted to a neighborhood around it.

Theorem 2 (Informal)**.** *Let* $\mathcal{A}$ *correspond to matrix factorization, sensing or completion under the assumptions in Section 1.2 and* $\beta = \Omega(poly(d))$*. If initialized close to one of* $\mathbf{E}_i, i \in \{1, 2\}$*, after a polynomial number of steps the discretized Langevin dynamics will converge to a distribution that is close in total variation distance to* $p(X)$ *when restricted to a neighborhood of* $\mathbf{E}_i$*.*

By way of remarks: we are interested in the scenario when the distribution is reasonably concentrated around the manifolds $\mathbf{E}_i, i \in \{1, 2\}$ – so *some* dependence of β on d is necessary. Furthermore, for the problem to be statistically interesting, some dependence on d is also necessary: previous work by Perry et al. (2018) (and a precursor by Péché (2006)) show that for certain priors over X (a particularly natural one is the spiked Wigner prior, where the ground truth matrix is rank-1, with random ± 1 entries), when $\beta < \frac{1}{d}$, no statistical test can distinguish the "planted" distribution from Gaussian nose with probability $o(1)$.

Note that importantly, all of our algorithms are **not** given an explicit description of the manifold around which they want to sample. The manifold is implicitly defined through $f(X)$ and our algorithms only use query access to its value and gradients. Nevertheless Langevin dynamics is able to discover this manifold on its own regardless of how it is embedded. The fact that the Euclidean metric is not the most natural metric for the manifolds of interest make many of the differential geometric quantities (like Ricci curvature) that we need estimates for quite challenging to get a handle on.

[1]Note, the gradient is for a function defined on the manifold $\mathbf{B}$.

References

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.

Sébastien Bubeck, Ronen Eldan, and Joseph Lehec. Sampling from a log-concave distribution with projected langevin monte carlo. *arXiv preprint arXiv:1507.02564*, 2015.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017a.

Gintare Karolina Dziugaite and Daniel M Roy. Entropy-sgd optimizes the prior of a pac-bayes bound: Data-dependent pac-bayes priors via differential privacy. *arXiv preprint arXiv:1712.09376*, 2017b.

Rong Ge, Holden Lee, and Andrej Risteski. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin monte carlo. In *Advances in neural information processing systems*, 2018.

Sandrine Péché. The largest eigenvalue of small rank perturbations of hermitian random matrices. *Probability Theory and Related Fields*, 134(1):127–173, 2006.

Amelia Perry, Alexander S Wein, Afonso S Bandeira, Ankur Moitra, et al. Optimality and sub-optimality of pca i: Spiked random matrix models. *The Annals of Statistics*, 46(5):2416–2451, 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.