===== **PROBABILITY THEORY AND MATHEMATICAL STATISTICS** =====

Non-Asymptotic Guarantees for Sampling by Stochastic Gradient Descent

A. G. Karagulyan[1*]

[1]*Yerevan State University, Yerevan, Armenia*
Received November 2, 2018; Revised December 25, 2018; Accepted January 24, 2019

Abstract—Sampling from various kinds of distributions is an issue of paramount importance in statistics since it is often the key ingredient for constructing estimators, test procedures or confidence intervals. In many situations, the exact sampling from a given distribution is impossible or computationally expensive and, therefore, one needs to resort to approximate sampling strategies. However, it is only very recently that a mathematical theory providing non-asymptotic guarantees for approximate sampling problem in the high-dimensional settings started to be developed. In this paper we introduce a new mathematical framework that helps to analyze the Stochastic Gradient Descent as a method of sampling, closely related to Langevin Monte-Carlo.

1. INTRODUCTION

Let us first introduce the mathematical setting of Langevin sampling. The general problem is to sample from the log-concave distribution with density $\pi(\theta) = c\exp(-f(\theta))$, where $f : \mathbb{R}^p \to \mathbb{R}$ satisfies the following two conditions:

$$\text{Strong convexity}: f(\theta_2) \geq f(\theta_1) + \nabla f(\theta_1)^T(\theta_2 - \theta_1) + \frac{m}{2}\|\theta_1 - \theta_2\|_2^2; \tag{1.1}$$

$$\text{Smoothness}: \|\nabla f(\theta_1) - \nabla f(\theta_2)\|_2 \leq M\|\theta_1 - \theta_2\|_2, \tag{1.2}$$

for all p-dimensional real vectors θ_1 and θ_2. The parameters m and M are positive numbers and $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^p$. The problem of sampling from π is closely related to the problem of finding the minimum of the function $f : \mathbb{R}^p \to \mathbb{R}$. Indeed, suppose we manage to sample from the distribution $\pi_\beta(\theta) = c_\beta \exp(-\beta f(\theta))$, where β is a large positive number. Then π_β will mainly be concentrated around the unique minimum point of f and it will have some kind of a spike form. Thus, a sample from π_β is a high probability approximation of the minimum point. Therefore considering f to be convex will facilitate our task for characterizing the convergence of the considered sampling method. For more details see [5] and [10].

Langevin Monte-Carlo algorithm is one of the methods for the approximate sampling from the target distribution π. The idea comes from the following Stochastic Differential Equation (SDE), named Langevin diffusion:

$$dX(t) = -\nabla f(X(t))dt + \sqrt{2}dW(t). \tag{1.3}$$

Here W is the standard Wiener process or Brownian motion in $\mathbb{R}^p$. The solution of (1.3) is a Markov process having π as invariant distribution [1]. In order to use this fact for our goal, we will use Euler-Maruyama discretization of (1.3), which can be found in [15]. It goes as follows:

$$\theta_{k+1} = \theta_k - h_{k+1}\nabla f(\theta_k) + \sqrt{2h_{k+1}}\xi_{k+1}, \tag{1.4}$$

*E-mail: avetik.karagulyan@ensae.fr

where $\xi_1, \xi_2, \ldots, \xi_k, \ldots$ follow Gaussian distribution $\mathcal{N}(0, I_p)$ and are independent from each other and θ_0. The latter is the starting point for the algorithm and it can be random as well. In particular when the step-sizes h_k are constantly equal to h and h is small, then for large enough k's the distance (Wasserstein, Total Variation) between the distribution of θ_k and π is small. This algorithm is called Gradient Langevin Dynamics (GLD) or Langevin Monte-Carlo (LMC) and it is actively studied nowadays ([3]–[9]).

In this paper, however we are not going to study the convergence of LMC algorithms. Instead we will review Stochastic Gradient Descent as a sampling method and represent it as a sampling algorithm. Let us recall SGD for the case of optimization. Often in Machine Learning problems we need to minimize the empirical risk. The latter is usually a sum-decomposable function

$$f : \mathbb{R}^p \mapsto \mathbb{R} : \quad f(x) = \sum_{i=1}^n g_i(x),$$

where n is the sample size and $g_i : \mathbb{R}^p \to \mathbb{R}$, for every $i = 1, \ldots, n$. The classical algorithm to solve a minimization problem, when mild assumptions are satisfied, is the Gradient Descent. Unfortunately when the sample size is large then every step of Gradient Descent is becoming computationally expensive. That is why Stochastic Gradient Descent is introduced. The main idea of SGD is to replace the full gradient in GD with its unbiased estimate. There are various ways to do it, but the most common one is the so called Batch Gradient Descent. In the latter case, one just samples a mini-batch B (a subset of $\{1, 2, \ldots, n\}$) and replaces the gradient by $c_B \sum_{i \in B} \nabla g_i$, where c_B is a constant depending on $|B|$. Thus the update rule becomes $\theta_{k+1} = \theta_k - c_B \sum_{i \in B} \nabla g_i$. For more details see [2].

The problem of our interest however is not directly related to optimization, but to sampling. We show that in the case of a smooth and strongly convex potential function f SGD yields a convergence of order $\tilde{O}(\kappa^2 p/\varepsilon^2)$ [1] in Wasserstein error. If in addition to these conditions we also have second-order smoothness, then the rate improves to $\tilde{O}(\kappa^2 p/\varepsilon^2 \bigwedge \kappa \sqrt{np}/\varepsilon)$.

This article is organized as follows: Section 2 gives some notions about the prior work in Langevin sampling. In Section 3 we introduce the theoretical setting that we are going to work with. In Section 4 we propose a mathematical framework which helps to analyze the convergence. The main results that provide non-asymptotic upper bounds to convergence rate are presented in Section 5.

2. PRIOR WORK

The first and probably the most influential work providing probabilistic analysis of the asymptotic properties of the LMC algorithm is decsribed in [15]. However, one of the recommendations made by the authors of that paper is to avoid using Langevin algorithm as it is defined in (1.4) or to use it very cautiously, since the ergodicity of the corresponding Markov chain θ_k is very sensitive to the choice of the parameter h. Even in the cases where the Langevin diffusion is geometrically ergodic, the inappropriate choice of h may result the transience of the Markov chain. These findings have strongly influenced the subsequent studies since all the ensuing research focused essentially on the Metropolis adjusted version of the LMC, known as Metropolis adjusted Langevin algorithm (MALA) and its numerous modifications ([11]–[16]). In contrast to this, it is shown that under the strong convexity assumption imposed on f coupled with the Lipschitz continuity of the gradient of f, one can ensure the non-transience of the Markov chain θ_k by a suitable choice of h_k. Later by [5] and [8] it was shown that the convergence rate in TV distance is $\tilde{O}(p/\varepsilon^2)$ for any initial vector θ_0.

Another problem of interest is the convergence in Wasserstein distance. In the next section the reader can find our reasoning to choose Wasserstein distance instead of TV. The convergence of LMC with this error was recently studied by [6] and [8] and a rate of $\tilde{O}(p/\varepsilon^2)$ was achieved. In addition to this, in [6] it was shown, that imposing additional smoothness for function f, meaning Lipschitz-continuity of its Hessian matrix, implies a better convergence rate of $\tilde{O}(\sqrt{p}/\varepsilon)$ for LMC. It turns out that in the case of sum-decomposable potential function, a modified version of LMC achives a better convergence rate.

---

[1] We denote the condition number M_g/m_g by κ, $\tilde{O}$ is the big-O notation, ignoring logarithmic factors.

Some of these algorithms have their roots in optimization, like SAGA [4], which was originally proposed in a paper by Defazio et al. [7] for the problem of optimization.

The convergence in terms of Wasserstein error was studied by many authors. [8] proved the rate $O(p/\varepsilon^2)$ for any deterministic starting point θ_0. The same convergence with improved coefficients was later shown in [6]. In this section we will formulate two theorems from [6], which will be used later on. Before we state the theorems, let us define W_2 Wasserstein distance. For two probability measures μ and ν defined on $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, W_2 distance is defined by

$$W_2(\mu, \nu) = \left\{ \inf_{\eta \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^p \times \mathbb{R}^p} \|\theta - \theta'\|_2^2 d\eta(\theta, \theta') \right\}^{1/2}, \tag{2.1}$$

where the infimum is taken with respect to all joint distributions η having μ and ν as marginal distributions. Let us compare this distance to total variation distance. If we have small Wasserstein for some μ and ν, then it implies that their first order moments are also close. This property does not hold for the total variation distance. As an example one can check that $\|\delta_\theta - \delta'_\theta\|_{TV} = \mathbf{1}_{\theta \neq \theta'}$, whereas $W_2(\delta_\theta, \delta'_\theta) = \|\theta - \theta'\|_2$ is a smooth function increasing function of Euclidean distance between θ and θ'.

Let us now present a non-asymptotic convergence bound for Wasserstein error, when the constant step-size LMC .

Theorem 2.1 (Theorem 1 from [6]). *Assume that $h \in (0, 2/M)$. Let f satisfy conditions* (1.1) *and* (1.2)*, thus the following claims hold:*

$$\text{if } h \leq \frac{2}{m + M}, \quad \text{then } W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{1.65M}{m}(hp)^{1/2};$$

$$\text{if } h \geq \frac{2}{m + M}, \quad \text{then } W_2(\nu_K, \pi) \leq (Mh - 1)^K W_2(\nu_0, \pi) + \frac{1.65Mh}{2 - Mh}(hp)^{1/2}.$$

In practice, a relevant approach to get an accuracy of at most ε is to minimize the upper bound provided by Theorem 2.1 with respect to h, for a fixed K. Then, one can choose the smallest K for which the obtained upper bound is smaller than ε. One useful observation is that the second upper bound is an increasing function of h. Its minimum is always attained at $h = 2/(m + M)$, which means that one can always look for a step-size in the interval $(0, 2/(m + M)]$ by minimizing the first upper bound. This can be done using standard methods of optimization.

Remark 2.1. *These two upper bounds contain $W_2(\nu_0, \pi)$, computation of which can be involving. In order to avoid it, we will bound it from above. If $f \geq 0$, we can replace it by $\sqrt{p/m} + \sqrt{2f(\theta_0)/m}$. Indeed,*

$$W_2(\nu_0, \pi) \leq \sqrt{\frac{p}{m}} + \|\theta_0 - \theta\|_2 \leq \sqrt{\frac{p}{m}} + \sqrt{\frac{2}{m}\big(f(\theta_0) - f(\theta_*)\big)} \leq \sqrt{\frac{p}{m}} + \sqrt{\frac{2f(\theta_0)}{m}}.$$

The first inequality is a corollary from Proposition 1 of [8]. Combining Theorem 2.1 with its remarks we obtain the following. Suppose that we choose h and K so that

$$h \leq \min\left(\frac{2}{m + M}, \frac{m^2\varepsilon^2}{11M^2p}\right) \text{ and } hK \geq \frac{1}{m}\log\left(Q(p, \varepsilon)\right), \tag{2.2}$$

where $Q(p, \varepsilon) = \frac{2f(\theta_0) + mp}{0.5m\varepsilon}$ is a real-valued rational function. Then each of the components from the right-hand side of Theorem will be less than 0.5ε, thus $W_2(\nu_K, \pi) \leq \varepsilon$.

Now, let us discuss the convergence rate of LMC in the case of additional smoothness. Below we present a theorem that quantifies the non-asymptotic behavior of LMC, when the potential function has a Lipschitz-continuous Hessian. That is, for every $x, y \in \mathbb{R}^p$ we have

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L\|x - y\|_2, \tag{2.3}$$

where $\| \cdot \|$ is the operator norm of matrices.

Theorem 2.2 (Theorem 4 from [6]). *Let ν_K be the distribution of K-th iterate of the LMC algorithm iterations. Assume that the function $f : \mathbb{R}^p \to \mathbb{R}$ satisfies* (1.1), (1.2) *and it is also L-Hessian-Lipschitz. Then for every $h < 2/(m + M)$,*

$$W_2(\nu_K, \pi) \leq (1 - mh)^K W_2(\nu_0, \pi) + \frac{Lhp}{2m} + \frac{11M^{\frac{3}{2}}h\sqrt{p}}{5m}.$$

Remark 2.2. *In order for the improvement of the rate to be visible, let us take a closer look to the order of step-size h and dimension p. Here we have $O(hp)$ meanwhile Theorem 2.1 gives only $O(\sqrt{hp})$, which is worse as h is considered to be small.*

Remark 2.3. *Doing analogous analysis as we did for the previous theorem, one can deduce that the convergence rate is $\tilde{O}(\sqrt{p}/\varepsilon)$.*

3. PROPOSED FRAMEWORK TO ANALYZE SGD

In the following sections we will discuss a special case for potential function f, in particular when f is a sum-decomposable function, that is

$$f(\theta) = \sum_{i=1}^{n} g(\theta, Z_i),$$

where n is a very large positive integer, $g : \mathbb{R}^p \times \mathcal{Z} \to \mathbb{R}$ is a given smooth function and $Z_1, \ldots, Z_n$ are iid random variables with values in some probability space $\mathcal{Z}$. To ease notation, we write $g_i(\theta) = g(\theta, Z_i)$. We assume here that the functions g_i are strongly convex with a coefficient m_g and its gradient is M_g Lipschitz-continuous. Therefore f is a convex and gradient-Lipschitz function as well, with coefficients nm_g and nM_g. So we have

$$\nabla f(\theta) = \sum_{i=1}^{n} \nabla g_i(\theta).$$

In order to avoid the computation of n gradients ∇g_i at each iteration of the LMC, we will use the classic Stochastic Gradient Descent algorithm in order to sample approximately. Let us first recall the algorithm. At each iteration k of the algorithm, we choose a subset B_k independent of all the past randomness and update θ_{k+1} by

$$\theta_{k+1} = \theta_k - \frac{hn}{b} \sum_{i \in B_k} \nabla g_i(\theta_k).$$

The latter can be rewritten as $\theta_{k+1} = \theta_k - h\nabla f(\theta_k) + h\zeta_k$, where the noise vectors ζ_k are of the form

$$\zeta_k = n\left\{ \frac{1}{b} \sum_{i \in B_k} \nabla g_i(\theta_k) - \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\theta_k) \right\}.$$

If b is large, the distribution of ζ_k (conditionally to θ_k) is approximately Gaussian $\mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma}_k)$ where the covariance matrix $\mathbf{\Sigma}_k$ is given by

$$\mathbf{\Sigma}_k = \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\theta_k) \nabla g_i(\theta_k)^\top - \left\{ \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\theta_k) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} \nabla g_i(\theta_k) \right\}^\top.$$

Below we study a particular case of SGD when the noise vector ζ_k is a normal random vector with a covariance proportional to identity matrix. We will assume, that $\mathbf{\Sigma}_k = \sigma^2 I_p$, where $\sigma^2 = n(n-b)/b$. The choice of σ^2 is intuitive. For details see the Appendix. Let us formulate the framework we are going to work with.

Assumptions: Suppose $g_i : \mathbb{R}^p \to \mathbb{R}^p$ for $i = 1, \ldots, n$ and $f = \sum_i g_i$. We will assume that the functions $g_1, g_2, \ldots, g_n$ satisfy the assumptions (1.1) and (1.2) with coefficients m_g and M_g, respectively.

Iterative method:

$$\theta_{k+1} = \theta_k - h\nabla f(\theta_k) + h\zeta_k, \tag{3.1}$$

where $\zeta_k \sim \mathcal{N}\left(0, \frac{n(n-b)}{b}I_p\right)$, for $k = 1, 2, \ldots, n$.

Problem: Find a solution to this optimization problem

$$\begin{cases} \textbf{Minimize } Kb; \\ \textbf{Subject to } \min_h W_2(\nu_{K,h,b}, \pi) \leq \varepsilon, \end{cases}$$

where $\nu_{K,h,b}$ is the distribution of the K-th iterate of the SGD with step-size h and batch-size b. In other words, what is the minimum amount of overall gradient evaluations in order to have an error of ε.

4. MAIN RESULTS

In this section we present two theorems that solve the problem stated above in two slightly different cases.

Theorem 4.1. *Suppose that the following conditions are satisfied:*

$$h = \frac{\varepsilon^2}{4\kappa^2 p}, \quad b = \frac{hn^2}{2 + hn}, \quad n \geq 9 \quad and \quad \frac{3\kappa\sqrt{p}}{n} \leq \varepsilon \leq \frac{2\kappa\sqrt{p}}{\sqrt{nM_g}}.$$

If

$$Kb \geq \frac{4p\kappa^2 n \log(Q'(p, \varepsilon))}{m_g(8p\kappa^2 + \varepsilon^2 n)}, \tag{4.1}$$

where Q' is a rational function given by formula

$$Q'(p, \varepsilon) = \frac{2f(\theta_0) + m_g p}{0.1 m_g \varepsilon},$$

then $W_2(\nu_{K,h,b}, \pi) \leq \varepsilon$.

Before we bring the proof let us state some remarks regarding this theorem.

Remark 4.1. *Since the batch-size b is between 1 and n, $hn^2/(2 + hn)$ must also satisfy this condition. In order to verify that, let us substitute h with its value. Therefore we have*

$$b = \frac{n^2 \varepsilon^2}{8\kappa^2 p + n\varepsilon^2}.$$

The latter is a monotonically increasing function with respect to ε^2. Thus taking into account that n is larger than 9,

$$b = \frac{n^2}{\frac{8\kappa^2 p}{\varepsilon^2} + n} \geq \frac{n^2}{\frac{8n^2}{9} + n} \geq 1. \tag{4.2}$$

The inequality $b \leq n$ is obvious.

Remark 4.2. *One can notice that, if $n \to \infty$, then Kb has an order of $\tilde{O}\left(\frac{4p\kappa^2}{\varepsilon^2}\right)$.*

Proof. As the function f is a sum of n strongly-convex and gradient-Lipschitz functions, then it is also a strongly-convex and gradient-Lipschitz function with coefficients $m = nm_g$ and $M = nM_g$, respectively. First let us express the step-size h in terms of the batch-size b. From the formula of b, we obtain $h = 2b/n(n - b)$. Thus if we can rewrite the iterative method in the following way:

$$\theta_{k+1} = \theta_k - h\nabla f(\theta_k) - h\zeta_k = \theta_k - h\nabla f(\theta_k) + h\sqrt{\frac{n(n-b)}{b}}\eta_k = \theta_k - h\nabla f(\theta_k) + \sqrt{2h}\eta_k,$$

where $\eta_1, \eta_2, \ldots$, as usual, are independent standard normal p-dimensional random vectors. Therefore we got the classic LMC update rule. From the definition of h we have

$$h = \frac{\varepsilon^2}{4\kappa^2 p} \le \frac{1}{nM_g}.$$

Thus Theorem 2.1 yields $W_2(\nu_K, \pi) \le (1 - nm_g h)^K Q(p, \varepsilon) + 1.65\kappa\sqrt{ph}$. We will give upper bounds for each component of the right-hand side. Substituting h with its value in $\kappa\sqrt{ph}$ we obtain, that

$$\kappa\sqrt{ph} = \kappa\sqrt{p\frac{\varepsilon^2}{4\kappa^2 p}} = \frac{\varepsilon}{2}.$$

Now let us discuss the other component. As we mentioned in previous sections, if

$$K \ge \frac{\log(Q'(p, \varepsilon))}{m_g nh} = \frac{4p\kappa^2}{m_g n\varepsilon^2} \cdot \log(Q'(p, \varepsilon)), \tag{4.3}$$

then $(1 - nm_g h)^K W_2(\nu_{0,h,b}, \pi)$ will be less than 0.1ε. In order to complete the proof we just need to multiply this lower bound on K by b. Thus we obtain

$$Kb \ge \frac{4p\kappa^2 b}{m_g n\varepsilon^2} \cdot \log(Q'(p, \varepsilon)). \tag{4.4}$$

Using the definition of h, we obtain the following formula for b

$$b = \frac{n^2\varepsilon^2}{8\kappa^2 p + n\varepsilon^2}.$$

Substituting the latter in (4.4), we get the required .

5. CONVERGENCE OF SGD WITH SECOND-ORDER SMOOTHNESS

In this section we analyze the convergence of Stochastic Gradient Descent in terms of Wasserstein distance when the Hessian matrix of the function f is Lipschitz-continuous.

Theorem 5.1. *Suppose that the following conditions are satisfied:*

$$h = \frac{\varepsilon}{4\kappa L_g\sqrt{M_g p \max(p, n)}}, \quad b = \frac{hn^2}{2 + hn},$$

$$\frac{2\sqrt{p \max(p, n)}}{n(n-1)} \le \frac{\varepsilon}{4\kappa L_g\sqrt{M_g}} \le \frac{\sqrt{p \max(p, n)}}{M_g n}.$$

If

$$Kb \ge \frac{4n\kappa L_g\sqrt{M_g p \max(p, n)}}{m_g(8\kappa L_g\sqrt{M_g p \max(p, n)} + n\varepsilon)} \cdot \log(Q''(p, \varepsilon)), \tag{5.1}$$

where

$$Q''(p, \varepsilon) = \frac{2f(\theta_0) + m_g p}{0.3 m_g \varepsilon},$$

then $W_2(\nu_{K,h,b}, \pi) \le \varepsilon$.

Remark 5.1. *Again the condition on ε is brought to make the choice of parameters possible. In particular, as mentioned before, b is an integer between 1 and n. Doing simple calculations and using the aforementioned condition, one can verify that our formula b satisfies this criteria.*

Remark 5.2. *Let us interpret a little the result of the theorem. In the case when our sample size n tends to infinity, we have $O\left(\kappa\sqrt{np}\log\left(Q(p,\varepsilon)\right)/\varepsilon\right)$ complexity.*

Proof. The proof is similar to the one for Theorem 4.1. Using the same reasoning as before f satisfies $(1.1),(1.2),(2.3)$ with $m = nm_g$, $M = nM_g$ and $L = nL_g$, respectively. As in the previous proof we will represent our iterative method as a classic Langevin Monte-Carlo update step. We have that

$$h = \frac{\varepsilon}{4\kappa L_g\sqrt{M_g p\max(p,n)}} \le \frac{1}{nM_g},$$

therefore 2.2 can be applied:

$$W_2(\nu_{K,h,b},\pi) \le (1 - nm_g h)^K W_2(\nu_{0,h,b},\pi) + \frac{L_g hp}{2m_g} + \frac{11}{5}\kappa h\sqrt{M_g pn}.$$

Let us express b in terms of ε, p and n:

$$b = \frac{hn^2}{2 + hn} = \frac{n^2}{\frac{2}{h} + n} = \frac{\varepsilon n^2}{8\kappa L_g\sqrt{M_g p\max(p,n)} + \varepsilon n}.$$

Thus the condition (5.1) is equivalent to

$$K \ge \frac{4\kappa L_g\sqrt{M_g p\max(p,n)}}{m_g n\varepsilon} \cdot \log\left(Q''(p,\varepsilon)\right) = \frac{\log\left(Q''(p,\varepsilon)\right)}{m_g nh}.$$

From the analysis shown above, this yields that $(1 - nm_g h)^K W_2(\nu_{0,h,b},\pi) \le 0.3\varepsilon$. Let us proceed to the second component, $L_g hp/2m_g$. From the formula of h, which is given in the statement of the theorem,

$$\frac{L_g hp}{2m_g} = \frac{L_g p}{2m_g} \cdot \frac{\varepsilon}{4\kappa L_g\sqrt{M_g p\max(p,n)}} \le \frac{\varepsilon}{8}.$$

The latter inequality is true, if we assume that L_g, M_g and κ are greater than 1. Similarly,

$$\frac{11M_g^{\frac{3}{2}} h\sqrt{pn}}{5m_g} = \frac{11M_g^{\frac{3}{2}}\sqrt{pn}}{5m_g} \cdot \frac{\varepsilon}{4\kappa L_g\sqrt{M_g p\max(p,n)}} \le \frac{11\varepsilon}{20}.$$

Summing up these three inequalities we obtain that, $W_2(\nu_{K,h,b},\pi) \le \varepsilon$. $\blacksquare$

6. APPENDIX: THE CHOICE OF THE NOISE VARIANCE

In this section we give a little insight on why and how we chose the distribution of the noise vectors in 3. Suppose we have a set of n numbers $A = \{a_1, a_2, \ldots, a_n\}$. A random variable X is designed in the following way. We take a uniformly random subset I of A with a fixed size b from the class C_b of all subsets of fixed size b. Afterwards we calculate the value of $\frac{n}{b}\sum_{i \in I} a_i$ and assign it to X. One can easily claim that $\mathbb{E}[X] = \sum_{i=1}^{n} a_i$ and therefore if we assume a_i's to be of the same order, then $\mathbb{E}[X] = O(n)$. Important detail to notice is that it does not depend on b. Unfortunately the order of the variance is not that easy to guess, so we will hereby calculate it.

Proposition 6.1. *Let us define the variance of X by $\mathbb{V}ar[X]$. Then*

$$\mathbb{V}ar[X] = O\left(\frac{n(n-b)}{b}\right).$$

Proof.

$$\mathbb{V}ar[X] = \frac{1}{C_n^b}\sum_{I \in C_b}\left[\sum_{i \in I} a_i\right]^2 - \left(\sum_{i=1}^{n} a_i\right)^2 = \frac{n^2}{b^2 C_n^b}\sum_{I \in C_b}\left[\sum_{i \in I} a_i^2 + \sum_{i \ne j; i,j \in I} 2a_i a_j\right] - \left(\sum_{i=1}^{n} a_i\right)^2$$

$$=\frac{n^2 C_{n-1}^{b-1}}{b^2 C_n^b}\sum_{i=1}^{n} a_i^2 + \frac{n^2 C_{n-2}^{b-2}}{b^2 C_n^b}\sum_{i\neq j} a_i a_j - \sum_{i=1}^{n} a_i^2 - \sum_{i\neq j} a_i a_j = \frac{n-b}{b}\sum_{i=1}^{n} a_i^2 + \frac{b-n}{nb-b}\sum_{i\neq j} a_i a_j.$$

We know that $\sum_{i=1}^{n} a_i^2 = O(n)$ and $\sum_{i\neq j} 2a_i a_j = O\big(n(n-1)\big)$. Therefore the order of the variance is $O\left(n(n-b)/b\right)$.

Conclusion. In this paper we have introduced a new mathematical framework which helps to analyze Stochastic Gradient Descent as a sampling method, where the potential function is strongly convex and has a Lipschitz gradient. Considering the particular case, where the stochastic term is a normal random vector with a diagonal covariance matrix, we have shown a convergence rate of $\tilde{O}(p/\varepsilon^2)$. The latter is a massive improvement compared to the classic LMC which was giving only $\tilde{O}(np/\varepsilon^2)$. In the case when we also assumed second-order smoothness, we have got $\tilde{O}(p/\varepsilon^2 \bigwedge \kappa\sqrt{np}/\varepsilon)$ convergence rate.

REFERENCES

1. D. Bakry, P. Cattiaux, and A. Guillin, "Rate of convergence for ergodic continuous Markov processes: Lyapunov versus Poincare", Journal of Functional Analysis, **254** (3), 727-759, 2008.
2. L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning", SIAM Review, **60** (2), 223-311, 2018.
3. N. Brosse et al. "Sampling from a log-concave distribution with compact support with proximal Langevin Monte Carlo", Proceedings of Machine Learning Research, **65**, 319-342, 2017.
4. N. S. Chatterji et al. "On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo", In: arXiv preprint arXiv:1802.05431, 2018.
5. 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), 651-676, 2017.
6. A. S. Dalalyan, A. G. Karagulyan, "User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient", arXiv preprint arXiv:1710.00095, 2017.
7. A. Defazio, F. R. Bach, and S. Lacoste-Julien, "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives", CoRR abs/1407.0202, arXiv: 1407.0202, 2014.
8. A. Durmus and E. Moulines, "High-dimensional Bayesian inference via the Unadjusted Langevin Algorithm", arXiv preprint arXiv:1605.01559, 2016.
9. A. Durmus and E. Moulines et al., "Nonasymptotic convergence analysis for the unadjusted Langevin algorithm", The Annals of Applied Probability, **27** (3), 1551-1587, 2017.
10. S. B. Gelfand and S. K. Mitter, "Recursive stochastic algorithms for global optimization in $\mathbb{R}^d$", SIAM Journal on Control and Optimization, **29** (5), 999-1018, 1991.
11. S. F. Jarner and E. Hansen, "Geometric ergodicity of Metropolis algorithms", Stochastic Process. Appl., **85** (2), 341-361, 2000.
12. N. S. Pillai, A. M. Stuart, and A. H. Thiery, "Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions", Ann. Appl. Probab., **22** (6), 2320-2356, 2012.
13. G. O. Roberts and J. S. Rosenthal, "Optimal scaling of discrete approximations to Langevin diffusions", J. R. Stat. Soc. Ser. B, Stat. Methodol., **60** (1), 255-268, 1998.
14. G. O. Roberts and O. Stramer, "Langevin diffusions and Metropolis-Hastings algorithms", Methodol. Comput. Appl. Probab., **4** (4), 337-357, 2003.
15. G. O. Roberts, R. L. Tweedie et al, "Exponential convergence of Langevin distributions and their discrete approximations", Bernoulli, **2** (4), 341-363, 1996.
16. O. Stramer and R. L. Tweedie, "Langevin-type models. I. Diffusions with given stationary distributions and their discretizations", Methodol. Comput. Appl. Probab., **1** (3), 283-306, 1999.