Biased Stochastic Gradient Methods for Log-sos-concave Desity Estimation

Wei Hu

August 11, 2014

1 Introduction

In this note, we present the analysis of convergence rates of biased stochastic gradient methods that could be used to calculate our log-sos-concave density estimator.

Recall that our objective function to minimize is

$$g(\theta) = \frac{1}{n}\theta^T \sum_{i=1}^{n} v_d(x_i) + \log \int_K e^{-\theta^T v_d(x)} dx, \qquad \theta \in \Theta_d \tag{1.1}$$

where Θ_d is the set of coeffients $\theta \in \mathbb{R}^s (s = \binom{p+d}{d})$ such that $\theta^T v_d(x)$ is an sos-convex polynomial, which is a closed convex set. The gradient and Hessian of g are

$$\begin{aligned} \nabla g(\theta) &= \frac{1}{n}\sum_{i=1}^{n} v_d(x_i) - \mathbb{E}_\theta(v_d(X)), \\ \nabla^2 g(\theta) &= \mathbb{V}_\theta(v_d(X)), \end{aligned}$$

where the expectation and covariance are taken with respect to the distribution $p_\theta(x) \propto e^{-\theta^T v_d(x)}$ whose support is a bounded set $K \subseteq [-R, R]^p$ for some $R > 0$. Since $\nabla^2 g(\theta) \succeq 0, \forall \theta$, g is a convex function.

2 Algorithm and Assumptions

Consider a convex optimization problem

$$\begin{aligned} \text{minimize} \quad & g(\theta) \\ \text{s.t.} \quad & \theta \in \Theta. \end{aligned} \tag{2.1}$$

The (projected) stochastic gradient method generates a sequence of feasible points $\{\theta_k\}_{k\geq 1}$ through the iterations

$$\theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_k(\nabla g(\theta_k) + \xi_k)), \qquad k = 1, 2, \cdots \tag{2.2}$$

where Π_Θ is the projection operator onto Θ, $\{\alpha_k\}$ is a deterministic positive sequence of step sizes, and ξ_k is the error in the evaluation of the gradient.

For our problem, the projection can be formulated as a semi-definite program, which can be solved efficiently. The evaluation of the gradient is through sampling, which is efficient for log-concave distribution.

The following is a list of assumptions that we will use: (we use $\|\cdot\|$ for 2-norm)

(H1) (2.1) has an optimal solution θ^*.

(H2) g is Lipschitz continuous, i.e., $\|\nabla g(\theta)\| \le M, \forall \theta$.

(H3) The error does not change the Lipschitz continuity of g, i.e., $\|\nabla g(\theta_k) + \xi_k\| \le M, \forall k$.

(H4) ∇g is Lipschitz continuous, i.e., $\|\nabla g(\theta) - \nabla g(\eta)\| \le L\|\theta - \eta\|, \forall \theta, \eta$.

(H5) Each visited point is at a bounded distance from the optimal point, i.e., $\|\theta_k - \theta^*\| \le D, \forall k$.

For our specific problem, we have the following proposition:

Proposition 1. *For our problem* (1.1)*, (H1), (H2), (H3) and (H4) are true.*

Proof. (1) Because Θ_d is a closed convex set and g is a convex function, $g(\theta)$ attains its minimum on Θ_d. (H1) is true.
(2) Since $K \in [-R, R]^p$, we have $v(x) \in [-R^d, R^d]^s$ for all $x \in K$. Hence $\frac{1}{n}\sum_{i=1}^n v_d(x_i) \in [-R^d, R^d]^s$ and $\mathbb{E}_\theta(v_d(X)) \in [-R^d, R^d]^s$, which implies

$$\nabla g(\theta) = \frac{1}{n}\sum_{i=1}^n v_d(x_i) - \mathbb{E}_\theta(v_d(X)) \in [-2R^d, 2R^d]^s.$$

Therefore

$$\|\nabla g(\theta)\| \le \sqrt{s(2R^d)^2} = 2R^d\sqrt{s}.$$

Let $M = 2R^d\sqrt{s} = 2R^d\sqrt{\binom{p+d}{d}}$, then (H2) is true.
(3) Suppose we estimate the gradient $\nabla g(\theta_k)$ by samples $x^{(1)}, x^{(2)}, \cdots, x^{(t)} \in K$, then their average $\frac{1}{t}\sum_{i=1}^t v(x^i) \in [-R^d, R^d]^s$ and

$$\nabla g(\theta_k) + \xi_k = \frac{1}{n}\sum_{i=1}^n v_d(x_i) - \frac{1}{t}\sum_{i=1}^t v(x^i) \in [-2R^d, 2R^d]^s.$$

Same as (2), we know that (H3) is true.
(4) To prove (H4), it suffices to prove that every eigenvalue of $\nabla^2 g(\theta)$ is at most L.

Each entry of $\nabla^2 g(\theta)$ has the form $\mathbb{E}_\theta(X^{\alpha+\beta}) - E_\theta(X^\alpha)E_\theta(X^\beta)$, where α and β are the multi-indices of monomials in p variables of degree at most d. So we have $\mathbb{E}_\theta(X^{\alpha+\beta}) - E_\theta(X^\alpha)E_\theta(X^\beta) \in [-2R^{2d}, 2R^{2d}]$.

By Gershgorin circle theorem, for any eigenvalue λ of $\nabla^2 g(\theta)$, there exists some row α such that

$$|\lambda - (\nabla^2 g(\theta))_{\alpha,\alpha}| \leq \sum_{\beta \neq \alpha} |(\nabla^2 g(\theta))_{\alpha,\beta}|,$$

which implies

$$\lambda \leq \sum_{\beta} |(\nabla^2 g(\theta))_{\alpha,\beta}| \leq s \cdot 2R^{2d}.$$

Let $L = s \cdot 2R^{2d} = 2R^{2d}\binom{p+d}{d}$, then (H4) is true.

□

Remark. Although we are unable to verify the correctness of (H5) for our problem, my idea is that we can assume the existence of some D such that (H5) is true.

3 Convergence Rate for Deterministic Errors

In this section, we regard the errors ξ_k are deterministic vectors and give convergence rate analyses for two types of step size choices. The first method is similar in [1] and relies on (H1), (H2), (H3) and (H5); the second is directly from [2] and relies on (H1), (H2), (H3) and (H4).

The results are summarized as the following theorems.

Theorem 2. *Assume (H1), (H2), (H3) and (H5), then we have*

$$g(\tilde{\theta}_N) - g(\theta^*) \leq \frac{\frac{1}{2}D^2 + D\sum_{k=1}^N \alpha_k \|\xi_k\| + \frac{1}{2}M^2 \sum_{k=1}^N \alpha_k^2}{\sum_{k=1}^N \alpha_k}. \tag{3.1}$$

Further, if the sequence of step sizes $\{\alpha_k\}$ is decreasing, then

$$g(\bar{\theta}_N) - g(\theta^*) \leq \frac{D^2}{2N\alpha_N} + D\frac{\sum_{k=1}^N \|\xi_k\|}{N} + \frac{M^2}{2N}\sum_{k=1}^N \alpha_k \tag{3.2}$$

where $\tilde{\theta}_N = \frac{\sum_{k=1}^N \alpha_k \theta_k}{\sum_{k=1}^N \alpha_k}$ and $\bar{\theta}_N = \frac{\sum_{k=1}^N \theta_k}{N}$ are the weighted and unweighted average of all visited points.

Proof. We know that the projection operator is non-expansive, i.e.,

$$\|\Pi_\Theta(\theta) - \Pi_\Theta(\eta)\| \leq \|\theta - \eta\|, \qquad \forall \theta, \eta.$$

Then we have

$$\begin{aligned}
\|\theta_{k+1} - \theta^*\|^2 &= \|\Pi_\Theta(\theta_k - \alpha_k(\nabla g(\theta_k) + \xi_k))) - \theta^*\|^2 \\
&= \|\Pi_\Theta(\theta_k - \alpha_k(\nabla g(\theta_k) + \xi_k))) - \Pi_\Theta(\theta^*)\|^2 \\
&\leq \|\theta_k - \alpha_k(\nabla g(\theta_k) + \xi_k) - \theta^*\|^2 \\
&= \|\theta_k - \theta^*\|^2 - 2\alpha_k(\nabla g(\theta_k) + \xi_k)^T(\theta_k - \theta^*) + \alpha_k^2\|\nabla g(\theta_k) + \xi_k\|^2 \\
&= \|\theta_k - \theta^*\|^2 - 2\alpha_k \nabla g(\theta_k)^T(\theta_k - \theta^*) - 2\alpha_k \xi_k^T(\theta_k - \theta^*) + \alpha_k^2\|\nabla g(\theta_k) + \xi_k\|^2.
\end{aligned} \tag{3.3}$$

By the convexity of g, we have $\nabla g(\theta_k)^T(\theta_k-\theta^*) \geq g(\theta_k)-g(\theta^*)$. By Cauchy-Schwarz inequality and (H5), we have $-\xi_k^T(\theta_k-\theta^*) \leq D\|\xi_k\|$. By (H3) we have $\|\nabla g(\theta_k)+\xi_k\|^2 \leq M^2$. Putting them together, we obtain

$$\|\theta_{k+1}-\theta^*\|^2 \leq \|\theta_k-\theta^*\|^2 - 2\alpha_k(g(\theta_k)-g(\theta^*)) + 2\alpha_k D\|\xi_k\| + \alpha_k^2 M^2 \tag{3.4}$$

or equivalently

$$\alpha_k(g(\theta_k)-g(\theta^*)) \leq \frac{1}{2}(\|\theta_k-\theta^*\|^2 - \|\theta_{k+1}-\theta^*\|^2) + \alpha_k D\|\xi_k\| + \frac{1}{2}\alpha_k^2 M^2. \tag{3.5}$$

(I) Proof of (3.1)

Sum (3.5) for $k = 1, \cdots, N$, then we get

$$\begin{aligned}\sum_{k=1}^N \alpha_k(g(\theta_k)-g(\theta^*)) &\leq \frac{1}{2}(\|\theta_1-\theta^*\|^2 - \|\theta_{N+1}-\theta^*\|^2) + D\sum_{k=1}^N \alpha_k\|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k^2 \\ &\leq \frac{1}{2}D^2 + D\sum_{k=1}^N \alpha_k\|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k^2.\end{aligned}$$

Therefore

$$\begin{aligned}g(\tilde{\theta}_N)-g(\theta^*) = g(\frac{\sum_{k=1}^N \alpha_k\theta_k}{\sum_{k=1}^N \alpha_k}) - g(\theta^*) &\leq \frac{\sum_{k=1}^N \alpha_k g(\theta_k)}{\sum_{k=1}^N \alpha_k} - g(\theta^*) = \frac{\sum_{k=1}^N \alpha_k(g(\theta_k)-g(\theta^*))}{\sum_{k=1}^N \alpha_k} \\ &\leq \frac{\frac{1}{2}D^2 + D\sum_{k=1}^N \alpha_k\|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k^2}{\sum_{k=1}^N \alpha_k}.\end{aligned} \tag{3.6}$$

(II) Proof of (3.2)

From (3.5) we have

$$g(\theta_k)-g(\theta^*) \leq \frac{1}{2\alpha_k}(\|\theta_k-\theta^*\|^2 - \|\theta_{k+1}-\theta^*\|^2) + D\|\xi_k\| + \frac{1}{2}\alpha_k M^2. \tag{3.7}$$

Sum (3.7) for $k = 1, \cdots, N$, then we get (note that $\{\alpha_k\}$ is decreasing)

$$\begin{aligned}\sum_{k=1}^N (g(\theta_k)-g(\theta^*)) &\leq \sum_{k=1}^N \frac{1}{2\alpha_k}(\|\theta_k-\theta^*\|^2 - \|\theta_{k+1}-\theta^*\|^2) + D\sum_{k=1}^N \|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k \\ &= \frac{\|\theta_1-\theta^*\|^2}{2\alpha_1} + \sum_{k=2}^N(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}})\|\theta_k-\theta^*\|^2 - \frac{\|\theta_{N+1}-\theta^*\|^2}{2\alpha_N} + D\sum_{k=1}^N \|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k \\ &\leq \frac{D^2}{2\alpha_1} + \sum_{k=2}^N(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}})D^2 + D\sum_{k=1}^N \|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k \\ &= \frac{D^2}{2\alpha_N} + D\sum_{k=1}^N \|\xi_k\| + \frac{1}{2}M^2\sum_{k=1}^N \alpha_k.\end{aligned}$$

Therefore

$$g(\bar{\theta}_N) - g(\theta^*) = g(\frac{\sum_{k=1}^N \theta_k}{N}) - g(\theta^*) \leq \frac{\sum_{k=1}^N g(\theta_k)}{N} - g(\theta^*) = \frac{\sum_{k=1}^N (g(\theta_k) - g(\theta^*))}{N}$$
$$\leq \frac{D^2}{2N\alpha_N} + D\frac{\sum_{k=1}^N \|\xi_k\|}{N} + \frac{M^2}{2N}\sum_{k=1}^N \alpha_k. \tag{3.8}$$

□

Corollary 3. *Assume (H1), (H2), (H3) and (H5). For step sizes* $\alpha_k = \frac{\beta}{Mk^r} (k = 1, \cdots, N)$ *for some* $0 < r < 1, \beta > 0$, *we have*

$$g(\tilde{\theta}_N) - g(\theta^*) \leq \frac{MD^2}{2\beta\sum_{k=1}^N k^{-r}} + D\frac{\sum_{k=1}^N k^{-r}\|\xi_k\|}{\sum_{k=1}^N k^{-r}} + \frac{\beta M\sum_{k=1}^N k^{-2r}}{2\sum_{k=1}^N k^{-r}} \tag{3.9}$$

and

$$g(\bar{\theta}_N) - g(\theta^*) \leq \frac{MD^2}{2\beta N^{1-r}} + D\frac{\sum_{k=1}^N \|\xi_k\|}{N} + \frac{\beta M\sum_{k=1}^N k^{-r}}{2N} \tag{3.10}$$

where $\tilde{\theta}_N = \frac{\sum_{k=1}^N \alpha_k\theta_k}{\sum_{k=1}^N \alpha_k}$ *and* $\bar{\theta}_N = \frac{\sum_{k=1}^N \theta_k}{N}$ *are the weighted and unweighted average of all visited points.*

Proof. Plugging $\alpha_k = \frac{\beta}{Mk^r}$ into (3.1) and (3.2), we get (3.9) and (3.10) respectively. □

Remark. If there is no error, i.e., $\xi_k = 0 (k = 1, 2, \cdots)$, then both (3.1) and (3.2) achieve optimal asymptotic bound $O(\frac{1}{\sqrt{N}})$ when $r = \frac{1}{2}$.

The following two theorems for basic and accelarated proximal-gradient methods are from [2]. Since projection is a special case of proximity operator, this method can be applied. The proofs of them are in [2].

Theorem 4. *(Basic proximal-gradient method) Assume (H1), (H2), (H3) and (H4). For step sizes* $\alpha_k = \frac{1}{L} (k = 1, \cdots, N)$, *we have*

$$g(\bar{\theta}_{N\setminus 1}) - g(\theta^*) \leq \frac{L}{2(N-1)}\big(\|\theta_1 - \theta^*\| + \frac{2}{L}\sum_{k=2}^N \|\xi_k\|\big)^2 \tag{3.11}$$

where $\bar{\theta}_{N\setminus 1} = \frac{\sum_{k=2}^N \theta_k}{N-1}$ *is the average of all visited points except* θ_1.

Theorem 5. *(Accelerated proximal-gradient method) Assume (H1), (H2), (H3) and (H4). If the algorithm starts from* $\theta_1 = \eta_1$, *and iterates as*

$$\begin{cases} \theta_{k+1} = \Pi_\Theta\big(\eta_k - \frac{1}{L}(\nabla g(y_k) + \xi_k)\big) \\ \eta_{k+1} = \theta_{k+1} + \frac{k-1}{k+2}(\theta_{k+1} - \theta_k), k = 1, 2, \cdots \end{cases} \tag{3.12}$$

then we have

$$g(\theta_N) - g(\theta^*) \leq \frac{2L}{N^2}\big(\|\theta_1 - \theta^*\| + \frac{2}{L}\sum_{k=2}^{N} k\|\xi_k\|\big)^2. \tag{3.13}$$

Remark. If there is no error, i.e., $\xi_k = 0 (k = 1, 2, \cdots)$, then the convergence rates of basic and accelerated proximal-gradient methods are $O(\frac{1}{N})$ and $O(\frac{1}{N^2})$ respectively.

4 Convergence Rate of Expectation for Stochastic Errors

In this section we will generalize the existing result for unbiased estimation of gradients to the biased case, which is the case if we use MCMC sampling to estimate the gradient.

We make the following assumption:

(H6) There exists an increasing sequences of σ-fields $\{\mathcal{F}_k\}_{k\geq 0}$ such that θ_k is $\mathcal{F}_{k-1}$-mesurable and $\|\mathbb{E}(\xi_k|\mathcal{F}_{k-1})\| \leq e_k$. $(k = 1, 2, \cdots)$

Note that for the unbiased case, we have $\epsilon_k = 0$. $(k = 1, 2, \cdots)$

Theorem 6. *Assume (H1), (H2), (H3), (H5) and (H6), then we have*

$$\mathbb{E}(g(\tilde{\theta}_N) - g(\theta^*)) \leq \frac{\frac{1}{2}D^2 + D\sum_{k=1}^{N}\alpha_k e_k + \frac{1}{2}M^2\sum_{k=1}^{N}\alpha_k^2}{\sum_{k=1}^{N}\alpha_k}. \tag{4.1}$$

Further, if the sequence of step sizes $\{\alpha_k\}$ is decreasing, then

$$\mathbb{E}(g(\bar{\theta}_N) - g(\theta^*)) \leq \frac{D^2}{2N\alpha_N} + D\frac{\sum_{k=1}^{N} e_k}{N} + \frac{M^2}{2N}\sum_{k=1}^{N}\alpha_k \tag{4.2}$$

where $\tilde{\theta}_N = \frac{\sum_{k=1}^{N}\alpha_k\theta_k}{\sum_{k=1}^{N}\alpha_k}$ and $\bar{\theta}_N = \frac{\sum_{k=1}^{N}\theta_k}{N}$ are the weighted and unweighted average of all visited points.

Proof. Recall that from (3.3) we have:

$$\begin{aligned}\|\theta_{k+1} - \theta^*\|^2 &\leq \|\theta_k - \theta^*\|^2 - 2\alpha_k\xi_k^T(\theta_k - \theta^*) + \alpha_k^2\|\nabla g(\theta_k) + \xi_k\|^2 \\ &\leq \|\theta_k - \theta^*\|^2 - 2\alpha_k(g(\theta_k) - g(\theta^*)) - 2\alpha_k\xi_k^T(\theta_k - \theta^*) + \alpha_k^2 M^2.\end{aligned}$$

Taking the expectation of the both sides, we obtain

$$2\alpha_k\mathbb{E}(g(\theta_k) - g(\theta^*)) \leq \mathbb{E}\|\theta_k - \theta^*\|^2 - \mathbb{E}\|\theta_{k+1} - \theta^*\|^2 - 2\alpha_k\mathbb{E}(\xi_k^T(\theta_k - \theta^*)) + \alpha_k^2 M^2. \tag{4.3}$$

From (H6) we have

$$\begin{aligned}\|\mathbb{E}(\xi_k^T(\theta_k-\theta^*))\| &= \|\mathbb{E}(\mathbb{E}(\xi_k^T(\theta_k-\theta^*)|\mathcal{F}_{k-1}))\| = \|\mathbb{E}\big(\mathbb{E}(\xi_k^T|\mathcal{F}_{k-1})(\theta_k-\theta^*)\big)\| \\ &\le \|\mathbb{E}(De_k)\| = De_k,\end{aligned}$$

then (4.3) implies

$$2\alpha_k\mathbb{E}(g(\theta_k)-g(\theta^*)) \le \mathbb{E}\|\theta_k-\theta^*\|^2 - \mathbb{E}\|\theta_{k+1}-\theta^*\|^2 + 2\alpha_k De_k + \alpha_k^2 M^2. \quad (4.4)$$

Then the proof goes exactly the same as the proof of Theorem 2, except that we use (4.4) in replace of (3.4). The proof is done. □

Remark. (4.1) is a generalization of (2.18.a) in [3], where unbiased estimation, i.e., $e_k = 0$, is assumed.

In our problem, if we use the sampling technique in [4] to estimate the gradients, we can obtain the bound e_k as follows. Following [4], suppose a sequence of log-concave distributions $\mu_1, \mu_2, \cdots$ is tracked, which correspond to our distributions $p_{\theta_1}, p_{\theta_2}, \cdots$. When the Markov chain is runned for sampling from μ_k, we can run the chain for enough steps to ensure the closeness in total variation distance from the current distribution $\hat{\mu}_k$ on the chain to μ_k, i.e.,

$$\int_K \big|d\hat{\mu}_k - d\mu_k\big| \le \varepsilon_k. \quad (4.5)$$

If we start taking samples $x^{(1)}, x^{(2)}, \cdots, x^{(t)}$ from now (assume that t is deterministic), where $x^{(i)}$ is taken from distribution $\hat{\mu}_k^{(i)}$ $(i = 1, 2, \cdots, t)$, then each $\hat{\mu}_k^{(i)}$ satisfies (4.5). So

$$\begin{aligned}\|\mathbb{E}(\xi_k|\mathcal{F}_{k-1})\| &= \|\mathbb{E}(\frac{1}{t}\sum_{i=1}^t v(x^{(i)})|\mathcal{F}_{k-1}) - \mathbb{E}_{\theta_k}(v(X))\| \\ &= \|\frac{1}{t}\sum_{i=1}^t \int_K v(x)d\hat{\mu}_k^{(i)} - \int_K v(x)d\mu_k\| \\ &\le \frac{1}{t}\sum_{i=1}^t \|\int_K v(x)(d\hat{\mu}_k^{(i)} - d\mu_k)\| \\ &\le \frac{1}{t}\sum_{i=1}^t \int_K \|v(x)\||d\hat{\mu}_k^{(i)} - d\mu_k| \\ &\le \frac{1}{t}\sum_{i=1}^t \int_K \sqrt{s(R^d)^2}|d\hat{\mu}_k^{(i)} - d\mu_k| \\ &= R^d\sqrt{s}\frac{1}{t}\sum_{i=1}^t \int_K |d\hat{\mu}_k^{(i)} - d\mu_k| \\ &\le R^d\sqrt{s}\varepsilon_k.\end{aligned} \quad (4.6)$$

Let $e_k = R^d\sqrt{s}\varepsilon_k = R^d\sqrt{\binom{p+d}{d}}\varepsilon_k$. We can make ε_k small enough to make e_k small.

References

[1] Honorio J. Convergence rates of biased stochastic optimization for learning sparse Ising models. *ICML*, 2012

[2] Schmidt, M., Le Roux, N., and Bach, F. Convergence rates of inexact proximal-gradient methods for convex optimization. *NIPS*, 2011.

[3] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.

[4] Narayanan, H., and Rakhlin, A. Efficient sampling from time-varying log-concave distributions. *aiXiv:1309.5977v1*, 2013.