# On Markov Chain Gradient Descent

Tao Sun (NUDT), Yuejiao Sun (UCLA), and Wotao Yin (UCLA)

nudtsuntao@163.com, sunyj@math.ucla.edu, wotaoyin@math.ucla.edu

## PROBLEM

$$\underset{x\in X\subseteq\mathbb{R}^n}{\text{minimize}}\ \mathbb{E}_\xi\left(F(x;\xi)\right) \tag{1}$$

where $\Pi$ is the distribution of a sample space $\Xi$, $X$ is closed and convex, $F(\cdot,\xi): X\to\mathbb{R}$ is convex or nonconvex but differentiable, associated with $\xi\in\Xi$.

## MOTIVATION

Instead of **Stochastic Gradient Descent (SGD)** (i.i.d sample $\xi^k$):

$$x^{k+1} = \mathbf{Proj}_X\left(x^k - \gamma_k \partial F(x^k;\xi^k)\right)$$

We use $\xi^k$ on a Markov-chain trajectory, and call this method **Markov Chain Gradient Descent (MCGD)**. Benefits:

- some distributions (e.g.,$\Xi := \{x\in\{0,1\}^n | \langle a,x\rangle \le b\}$) cannot be sampled directly, but they have Markov-chain samples.
- Markov chains naturally arise in some applications, e.g. linear dynamic systems with random transitions or errors, and distributed systems in which each node stores a subset of training samples.

## ADVANTAGE OF MCGD OVER SGD

We use a numerical example to illustrate the advantage of MCGD over SGD. Consider an auto regressive model:

$$\xi_t^1 = A\xi_{t-1}^1 + e_1 W_t,\ W_t \overset{\text{i.i.d}}{\sim} N(0,1)$$

$$\bar\xi_t^2 = \begin{cases} 1, & \text{if } \langle u,\xi_t^1\rangle > 0,\\ 0, & \text{otherwise;}\end{cases} \qquad \xi_t^2 = \begin{cases} \bar\xi_t^2, & \text{with probability 0.8,}\\ 1-\bar\xi_t^2, & \text{with probability 0.2.}\end{cases}$$

Clearly, $(\xi_t^1,\xi_t^2)_{t=1}^\infty$ forms a Markov chain. Let $\Pi$ denote the stationary distribution of this Markov chain. We recover $u$ as the solution to the following problem:

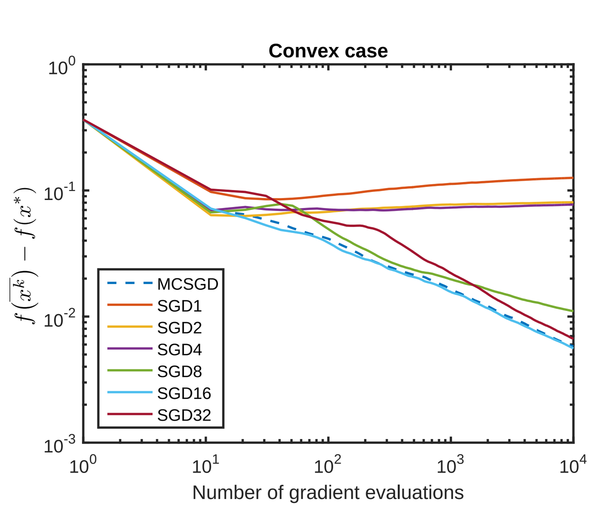
$$\underset{x}{\text{minimize}}\quad \mathbb{E}_{(\xi^1,\xi^2)\sim\Pi}\ell(x;\xi^1,\xi^2).$$

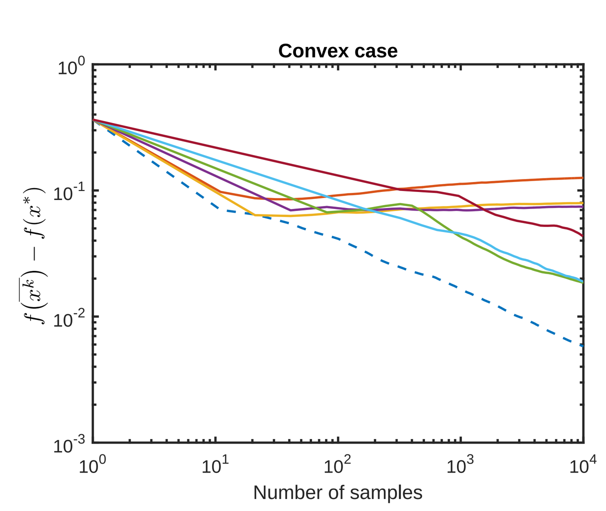
Compare:

1. MCGD, where samples are taken from one trajectory of the Markov chain;
2. SGD$T$, where $T$ is the burn-in time and each sample is the $T$th sample of a fresh, independent trajectory.

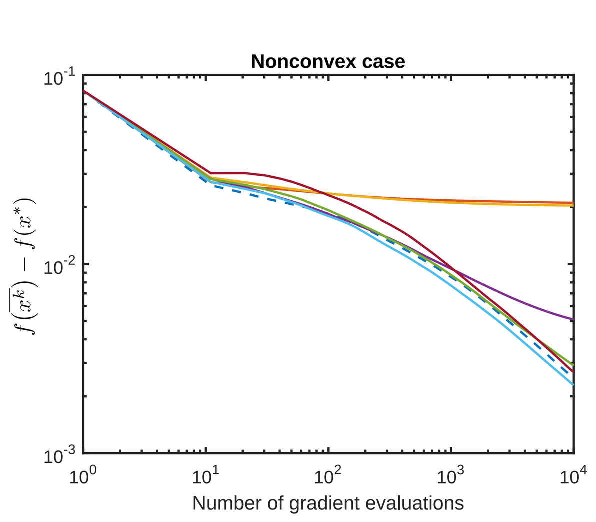
We test on both convex and nonconvex loss functions. ($\sigma(t) = \frac{1}{1+\exp(-t)}$).

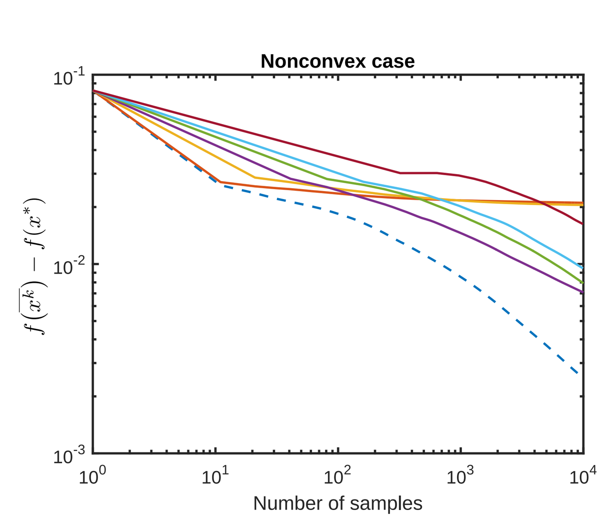
$$\ell(x;\xi^1,\xi^2) = -\xi^2\log(\sigma(\langle x,\xi^1\rangle)) - (1-\xi^2)\log(1-\sigma(\langle x,\xi^1\rangle)).$$




$$\ell(x;\xi^1,\xi^2) = \frac{1}{2}(\sigma(\langle x,\xi^1\rangle) - \xi^2)^2.$$




## CONTRIBUTIONS

All analyses of MCGD must deal with the biased expectation. Existing MCGD work [1-4]. New results of this work:

1. allow **non-reversible** Markov chain for faster convergence;
2. objective can be nonconvex.
3. non-ergodic convegence of the objective.

## CONVERGENCE ANALYSIS

### FINITE STATE SPACE

The analysis is based on the following assumptions:

1. The Markov chain $(X_k)_{k\ge0}$ is time-homogeneous, irreducible, and aperiodic. It has a transition matrix $P$ and stationary distribution $\pi^*$.
2. $X$ is convex and compact.
3. $\Xi$ is finite. Let $f_i(x) = M\cdot\mathbf{Prob}(\xi = y^i)\cdot F(x,y^i)$, and reformulate (1) as

$$\underset{x\in X\subseteq\mathbb{R}^d}{\text{minimize}}\ f(x) = \frac{1}{M}\sum_{i=1}^M f_i(x),$$

where each state $i$ has the uniform probability $1/M$. MCGD runs

$$x^{k+1} = \mathbf{Proj}_X(x^k - \gamma_k\partial f_{j_k}(x^k)),$$

where $(j_k)_{k\ge0}$ forms a Markov chain trajectory. It can be illustrated in the following diagram:

$$\begin{array}{ccccccccc} & & j_0 & \to & j_1 & \to & j_2 & \to & \dots \\ & & | & & | & & | & & \\ x^0 & \to & x^1 & \to & x^2 & \to & x^3 & \to & \dots \end{array}$$

#### CONVEX CASE

Assume that $f_i$, $i\in[M]$, are convex functions, and the stepsizes satisfy

$$\sum_k \gamma_k = +\infty,\quad \sum_k \ln k\cdot\gamma_k^2 < +\infty. \tag{2}$$

Let $\bar{x}^k = \frac{\sum_{i=1}^k \gamma_i x^i}{\sum_{i=1}^k\gamma_i}$, and we have

$$\lim_k \mathbb{E}f(x^k) = f^*,$$

$$\mathbb{E}\left(f(\bar{x}^k) - f^*\right) = O\left(\frac{\Phi(P)}{\sum_{i=1}^k\gamma_i}\right).$$

When choosing $\gamma_k = O(\frac{1}{k^q})$, $\frac{1}{2} < q < 1$,

$$\mathbb{E}\left(f(\bar{x}^k) - f^*\right) = O\left(\frac{\Phi(P)}{k^{1-q}}\right).$$

Note that the same stepsize and convergence rate can hold for SGD and subgradient algorithms.

#### NONCONVEX CASE

Assume that $X=\mathbb{R}^n$, $f_i$ is differentiable and $\nabla f_i$ is $L$-Lipschitz and bounded by $D>0$, and

$$\sum_k\gamma_k = +\infty,\ \sum_k \ln^2 k\cdot\gamma_k^2 < +\infty. \tag{3}$$

Let $f^* = \min_{x\in X} f(x)$. Then, we have

$$\lim_k \mathbb{E}\|\nabla f(x^k)\| = 0,$$

$$\mathbb{E}\left(\min_{1\le i\le k}\{\|\nabla f(x^k)\|^2\}\right) = O\left(\frac{\Phi(P)}{\sum_{i=1}^k\gamma_i}\right).$$

When choosing $\gamma_k = O(\frac{1}{k^q})$, $\frac{1}{2}<q<1$, the convergence rate is $O(\frac{\Phi(P)}{k^{1-q}})$.

### CONTINUOUS STATE SPACE

Assume that state space $\Xi$ is a continuum. Consider an infinite-state Markov chain that is time-homogeneous and reversible and solve (1) by MCGD.

#### CONVEX CASE

Assume that for each $\xi\in\Xi$, $F(\cdot;\xi)$ is convex, $|F(x;\xi) - F(y;\xi)| \le L\|x-y\|$, $\sup_{x\in X,\xi\in\Xi}\{\|\hat\nabla F(x;\xi)\|\} \le D$, $\mathbb{E}_\xi\hat\nabla F(x;\xi) \in \partial\mathbb{E}_\xi F(x;\xi)$, and $\sup_{x,y\in X,\xi\in\Xi}|F(x;\xi) - F(y;\xi)| \le H$. Choose $\gamma^k$ according to (2). Let$F^* := \min_{x\in X}\mathbb{E}_\xi(F(x;\xi))$. $\lambda\in(0,1)$ is the geometric rate of the mixing time of the Markov chain. Then we have

$$\lim_k \mathbb{E}\left(\mathbb{E}_\xi\left(F(x^k;\xi)\right) - F^*\right) = 0,$$

$$\mathbb{E}\left(\mathbb{E}_\xi\left(F(\bar{x}^k;\xi)\right) - F^*\right) = O\left(\frac{\max\{1,\frac{1}{\ln(1/\lambda)}\}}{\sum_{i=1}^k\gamma_i}\right),$$

#### NONCONVEX CASE

Let $X = \mathbb{R}^n$. Assume for any $\xi\in\Xi$, $F(x;\xi)$ is differentiable, and $\|\nabla F(x;\xi) - \nabla F(y;\xi)\| \le L\|x-y\|$. In addition, $\sup_{x\in X,\xi\in\Xi}\{\|\nabla F(x;\xi)\|\} < +\infty$, $X$ is the full space, and $\mathbb{E}_\xi\nabla F(x;\xi) = \nabla\mathbb{E}_\xi F(x;\xi)$. Then, we have

$$\lim_k \mathbb{E}\|\nabla\mathbb{E}_\xi(F(x^k;\xi))\| = 0,$$

$$\mathbb{E}\left(\min_{1\le i\le k}\{\|\nabla\mathbb{E}_\xi\left(F(x^i;\xi)\right)\|^2\}\right).$$

## ACCELERATION DUE TO NON-REVERSIBILITY

Non-reversibility can accelerate the mixing process of Markov chains. The following experiment compares MCGD with reversible and non-reversible Markov chains over the same graph with 20 nodes. The objective is a least square problem with data distributed on the graph.

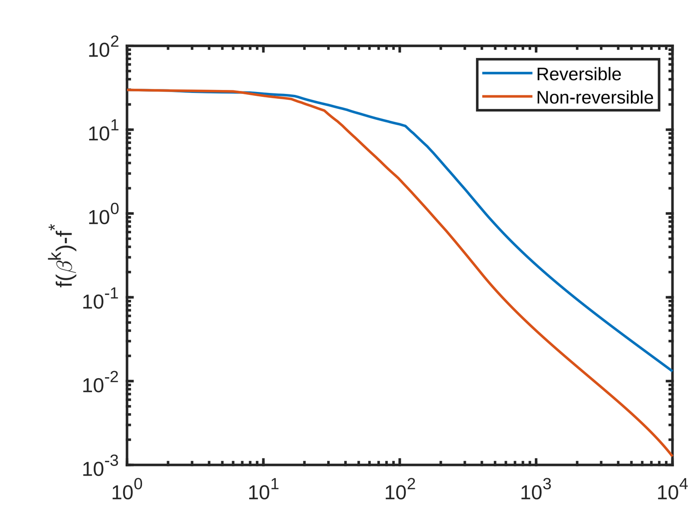



Figure 1: The second largest eigenvalues of reversible and non-reversible Markov chains are 0.75 and 0.66 respectively.

## REFERENCES

[1] JC Duchi, A Agarwal, M Johansson, MI Jordan. *Ergodic mirror descent*. SIAM Journal on Optimization, 2012.

[2] B Johansson, M Rabi, M Johansson. *A simple peer-to-peer algorithm for distributed optimization in sensor networks*. In Decision and Control, IEEE, 2007.

[3] B Johansson, M Rabi, M Johansson. *A randomized incremental subgradient method for distributed optimization in networked systems*. SIAM Journal on Optimization, 2009.

[4] SS Ram, A Nedic, VV Veeravalli. *Incremental stochastic subgradient algorithms for convex optimization*. SIAM Journal on Optimization, 2009.