On the Convergence of Monte Carlo Methods with Stochastic Gradients

Quanquan Gu

Department of Computer Science
UCLA

Simons Institute Workshop on Sampling Algorithms and Geometries on Probability Distributions

Joint work with Difan Zou and Pan Xu

10/1/2021

Sampling Problems

- The goal is to generate samples $\mathbf{x}$ from the probability density function $\pi(\mathrm{d}\mathbf{x})$.
- In many cases, the target distribution is represented by $\pi \propto e^{-f(\mathbf{x})}$, where the negative log-density function $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ is known and satisfies certain regularity conditions, i.e., (strongly) convex, smooth, etc.

Sampling Problems in Large-Scale Bayesian Learning

- In Bayesian Learning, the target distribution π is typically the posterior given i.i.d. observations $\{\mathbf{z}_i\}_{i=1,\ldots,n}$.

$$\pi = p(\mathbf{x} \mid \mathbf{z}_1, \ldots, \mathbf{z}_n) \propto p(\mathbf{z}_1, \ldots, \mathbf{z}_n \mid \mathbf{x}) \cdot p(\mathbf{x}) = p(\mathbf{x}) \cdot \prod_{i=1}^{n} p(\mathbf{z}_i \mid \mathbf{x})$$

Posterior **Likelihood** **Prior**

- Then π can be rewritten as

$$\pi \propto e^{-f(\mathbf{x})} = e^{-\sum_{i=1}^{n} f_i(\mathbf{x})} \quad \text{where} \quad f_i(\mathbf{x}) = -\log(p(\mathbf{z}_i \mid \mathbf{x})) - n^{-1} \cdot \log(p(\mathbf{x}))$$

Markov Chain Monte Carlo methods

- MCMC method
  - For $t = 1,\ldots,T$
    - **Proposal**: $\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{g}_f(\mathbf{x}_t)$ — A random vector depending on f and $\mathbf{x}_t$
    - **Reject**: $\mathbf{x}_{t+1} = \mathbf{x}_t$ with probability $1 - \alpha_f(\mathbf{x}_t, \mathbf{x}_{t+1})$

Metropolis-Hasting acceptance probability

- Examples: random walk Metropolis [Mengersen and Tweedie, 1996], ball walk [Lovasz and Simonovits, 1990], Metropolis-adjusted Langevin algorithms (MALA) [Robert and Tweedie 1996], Hamiltonian Monte Carlo (HMC) [Duane et. al., 1987]

Hamiltonian Monte Carlo

- ODE description Hamiltonian energy $H(\mathbf{x}, \mathbf{p}) = f(\mathbf{x}) + \|\mathbf{p}\|_2^2/2$

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \frac{\partial H(\mathbf{x}(t), \mathbf{p}(t))}{\partial \mathbf{p}} = \mathbf{p}(t) \qquad \frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = -\frac{\partial H(\mathbf{x}(t), \mathbf{p}(t))}{\partial \mathbf{x}} = -\nabla f(\mathbf{x}(t))$$

- (Idealized) Hamiltonian Monte Carlo Method
  - $\mathbf{x}_{t+1} = \mathbf{x}_t + \int_{\tau=0}^{\tau_0} \mathbf{p}(\tau)\mathrm{d}\tau$, where $\mathbf{x}(0) = \mathbf{x}_t$, $\mathbf{p}(0) \sim N(0, \mathbf{I})$

Key property: When $t \to \infty$, $\mathbf{x}_t \sim \pi \propto e^{-f(\mathbf{x})}$

Duane et. al., Hybrid monte carlo. Physics letters B, 1987

Underdamped Langevin Dynamics

- SDE description

Friction | Potential | Brownian motion

$$\mathrm{d}\mathbf{v}(t) = -\gamma\mathbf{v}(t)\mathrm{d}t - u\nabla f(\mathbf{x}(t))\mathrm{d}t + \sqrt{2\gamma u}\cdot \mathrm{d}\mathbf{B}(t)$$
$$\mathrm{d}\mathbf{x}(t) = \mathbf{v}(t)\mathrm{d}t$$

- (Idealized) Underdamped Langevin MCMC Method

  - $\mathbf{x}_{t+1} = \mathbf{x}_t + \int_{\tau=0}^{\eta} \mathbf{v}(\tau)\mathrm{d}\tau,$

    $\mathbf{v}_{t+1} = \mathbf{v}_t + \int_{\tau=0}^{\eta} -\left[\gamma\mathbf{v}(\tau) + u\nabla f(\mathbf{x}(\tau))\right]\mathrm{d}\tau + \sqrt{2\gamma u\eta}\cdot \boldsymbol{\xi}_t$

    where $\mathbf{v}(0) = \mathbf{v}_t$, $\mathbf{x}(0) = \mathbf{x}_t$, $\boldsymbol{\xi}_t \sim N(0,\mathbf{I})$

Key property: When $t \to \infty$, $(\mathbf{x}_t, \mathbf{v}_t) \sim \pi \propto e^{-f(\mathbf{x})-\|\mathbf{v}\|_2^2/2}$

Hendrik Anthony Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. Physica, 1940.

MCMC with Stochastic Gradients

- Both HMC and underdamped LMC involve the calculation of the gradient $\nabla f(\mathbf{x})$, which becomes inefficient when n is large.

- A commonly used solution is to calculate the stochastic gradient using a randomly sampled mini-batch of data.

HMC with Stochastic Gradients

- Stochastic Gradient Hamiltonian Monte Carlo Method
  - Input $\mathbf{x}_0, \eta, T, K$
  - For $t = 0, \ldots, T$
    - Let $\mathbf{p}_0 \sim \mathcal{N}(0, \mathbf{I})$
    - Let $\mathbf{q}_0 = \mathbf{x}_t$
    - For $k = 0, \ldots, K-1$
      - $\mathbf{p}_{k+1/2} = \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_k, \xi_k)$
      - $\mathbf{q}_{k+1} = \mathbf{q}_k + \eta \mathbf{p}_{k+1/2}$
      - $\mathbf{p}_{k+1} = \mathbf{p}_k - \frac{\eta}{2}\mathbf{g}(\mathbf{q}_{k+1}, \xi_{k+1/2})$
    - Let $\mathbf{x}_{t+1} = \mathbf{q}_K$ **Skip the MH step**
  - Output $\mathbf{x}_T$

Proposal: Numerically solving Hamilton's equation via stochastic gradients $\mathbf{g}(\mathbf{q}_k, \xi_k)$

Leapfrog numerical integrator

Zou and Gu, On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients, ICML 2021

Key Questions in the Convergence Analysis

- **Inner Loop:** What's the approximation error of the Leapfrog integrator using stochastic gradients?

- **Outer Loop:** Can the approximate ODE solutions lead to small sampling error?

Assumptions on the Target Distribution

- Assumptions:
  - Strongly log-concave distribution: $f(\mathbf{x})$ is μ-strongly convex
  - Log-smooth distribution: $f(\mathbf{x})$ is L-smooth,
  - Define $\kappa = L/\mu$ be the condition number
  - Bounded variance: For all iterate $\mathbf{q}_k$, $\mathbb{E}[\|\mathbf{g}(\mathbf{q}_k, \xi_k) - \nabla f(\mathbf{q}_k)\|_2^2] \leq \sigma^2$, where the expectation is taken on both $\mathbf{q}_k$ and ξ_k.

Approximation Error of the Numerical ODE Solver (Inner Loop)

- Define 3 sequences ($\mathbf{q}_0 = \mathbf{x}_t$):

$$(\mathcal{S}_\eta \mathbf{q}_k, \mathcal{S}_\eta \mathbf{p}_k) = (\mathbf{q}_{k+1}, \mathbf{p}_{k+1})$$

HMC with stochastic gradient

$$(\mathcal{G}_\eta \mathbf{q}_k, \mathcal{G}_\eta \mathbf{p}_k) = (\mathbb{E}[\mathbf{q}_{k+1} \,|\, \mathbf{p}_k, \mathbf{q}_k], \mathbb{E}[\mathbf{p}_{k+1} \,|\, \mathbf{p}_k, \mathbf{q}_k])$$

Conditionally expected stochastic gradient HMC update

$$(\mathcal{H}_\eta \mathbf{q}_k, \mathcal{H}_\eta \mathbf{p}_k) = \left(\mathbf{q}_k + \int_0^\eta \mathbf{p}(t)\mathrm{d}t, \mathbf{p}_k - \int_0^\eta \nabla f(\mathbf{q}(t))\mathrm{d}t\right)$$

Update via exact ODE solution

- Approximation error: we want to characterize the difference between $\mathcal{S}_\eta^K \mathbf{q}_0$ and $\mathcal{H}_\eta^K \mathbf{q}_0$.

Decomposition of the Approximation Error (Inner Loop)

- Define $\mathbf{z}_k = \begin{pmatrix} \mathbf{q}_k \\ L^{-1/2}\mathbf{p}_k \end{pmatrix} = \mathcal{S}_\eta^k \begin{pmatrix} \mathbf{q}_0 \\ L^{-1/2}\mathbf{p}_0 \end{pmatrix} = \mathcal{S}_\eta^k \mathbf{z}_0$, then

$$\mathcal{E}_k := \mathbb{E}\left[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right] = \mathbb{E}\left[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 + \mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right]$$
$$= \mathbb{E}\left[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0\|_2^2\right] + \mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right]$$

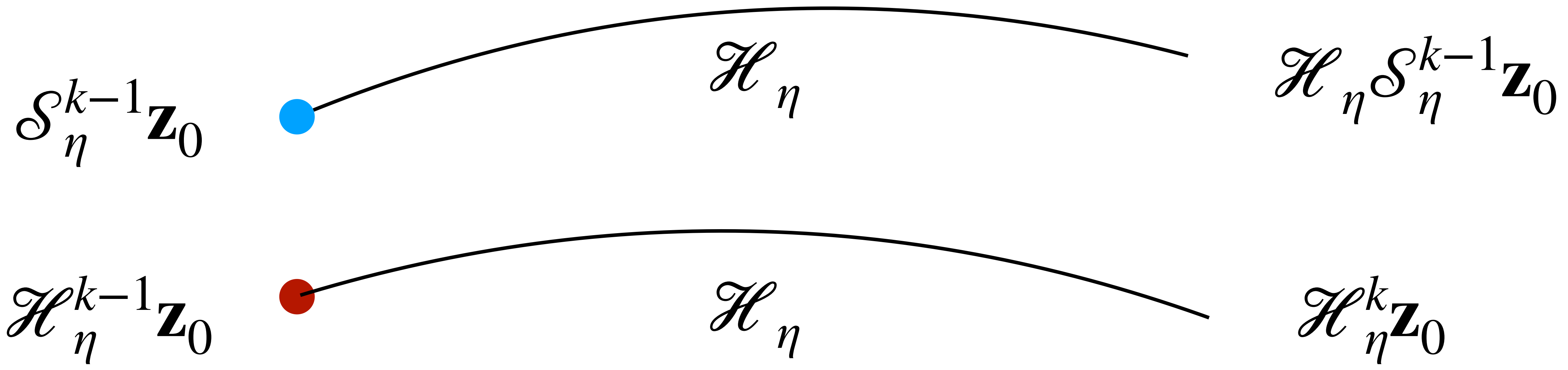
One-step statistical error between $\mathcal{S}_\eta$ and $\mathcal{G}_\eta$: $= O(L^{-1} \cdot \sigma^2 \cdot \eta^2)$

$$\mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right] = \mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 + \mathcal{H}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right]$$
$$\leq (1+\alpha) \cdot \mathbb{E}\left[\|\mathcal{H}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta^k \mathbf{z}_0\|_2^2\right]$$
$$+(1+1/\alpha) \cdot \mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0 - \mathcal{H}_\eta \mathcal{S}_\eta^{k-1}\mathbf{z}_0\|_2^2\right]$$

One-step "discretization error" between $\mathcal{G}_\eta$ and $\mathcal{H}_\eta$: $= O(Ld \cdot \eta^4)$

Decomposition of the Approximation Error (Inner Loop)

- Bound on $\mathbb{E}\left[\|\mathscr{H}_\eta \mathscr{S}_\eta^{k-1}\mathbf{z}_0 - \mathscr{H}_\eta^k \mathbf{z}_0\|_2^2\right]$



- $\mathscr{H}_\eta$ does not have contraction property on any two different points but has bounded expansion property

$$\mathbb{E}\left[\|\mathscr{H}_\eta \mathscr{S}_\eta^{k-1}\mathbf{z}_0 - \mathscr{H}_\eta^k \mathbf{z}_0\|_2^2\right] \leq e^{2L^{1/2}\eta} \cdot \mathbb{E}\left[\|\mathscr{S}_\eta^{k-1}\mathbf{z}_0 - \mathscr{H}_\eta^{k-1}\mathbf{z}_0\|_2^2\right] = e^{2L^{1/2}\eta} \cdot \mathscr{E}_{k-1}$$

Upper Bound of the Approximation Error

- Putting things together

Expansion term — One-step error

$$\mathscr{E}_k \leq (1+\alpha)\cdot e^{2L^{1/2}\eta}\cdot \mathscr{E}_{k-1} + (1+1/\alpha)\cdot O(Ld\cdot\eta^4) + O(L^{-1}\cdot\sigma^2\cdot\eta^2)$$

$$\leq \frac{e^{(2L^{1/2}\eta+\alpha)k}}{2L^{1/2}\eta+\alpha}\cdot\left[(1+1/\alpha)\cdot O(Ld\cdot\eta^4) + O(L^{-1}\cdot\sigma^2\cdot\eta^2)\right]$$

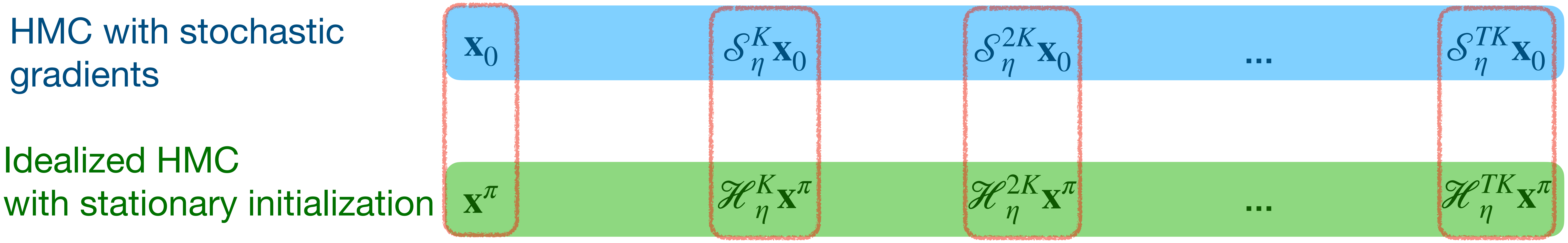
- Then we can set $\alpha = 2L^{1/2}\eta$ such that if $K\eta \leq 1/(4L^{1/2})$,

$$\mathscr{E}_K = \mathbb{E}\left[\|\mathscr{S}_\eta^K \mathbf{q}_0 - \mathscr{H}_\eta^K \mathbf{q}_0\|_2^2\right] \leq O(d\eta^2 + L^{-3/2}\cdot\sigma^2\cdot\eta)$$

Convergence Analysis of Outer Loop

- The key is to show that the approximation error will not explode.

- Analysis framework:



- Sampling error: we will characterize the difference between $\mathcal{S}_\eta^{TK}\mathbf{x}_0$ and $\mathcal{H}_\eta^{TK}\mathbf{x}^\pi$.

Contraction Property in the Outer Loop

- $\mathcal{H}_t$ has a good contraction property for any two points with the same velocity [Chen and Vempala19]: for any two points $(\mathbf{q}, \mathbf{p})$ and $(\mathbf{q}', \mathbf{p})$, then for any $0 \le t \le 1/(2\sqrt{L})$,

$$\mathbb{E}\left[\|\mathcal{H}_t\mathbf{q} - \mathcal{H}_t\mathbf{q}'\|_2^2\right] \le (1 - \mu t^2)\|\mathbf{q} - \mathbf{q}'\|_2^2$$

Strongly log-concave parameter

- Decomposition of the error propagation ($K\eta = 1/(4L^{1/2})$)

$$\mathbb{E}\left[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2\right] \le (1+\beta)\|\mathcal{H}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0'\|_2^2 + (1+1/\beta)\mathbb{E}\left[\|\mathcal{S}_\eta^K\mathbf{q}_0 - \mathcal{H}_\eta^K\mathbf{q}_0\|_2^2\right]$$

Contracting term

$$\le (1 - 1/(16\kappa))\|\mathbf{q}_0 - \mathbf{q}_0'\|_2^2$$

Approximation error

$$= O(d\eta^2 + L^{-3/2}\cdot\sigma^2\cdot\eta)$$

Setting $\beta = 1/(32\kappa)$ can avoid error explosion.

Chen and Vempala, Optimal convergence rate of Hamiltonian Monte Carlo for strongly log-concave distributions, APPROX-RANDOM 2019

Convergence Rates of Stochastic Gradient HMC

Theorem [Zou and Gu, 2021] Suppose all assumptions are satisfied, set $K = 1/(4\sqrt{L}\eta)$, then,

$$\mathscr{W}_2^2\left(\mathbf{P}(\mathbf{x}_T), \pi\right) \leq e^{-T/(32\kappa)} \cdot \mathbb{E}\left[\|\mathbf{x}_0 - \mathbf{x}^\pi\|_2^2\right] + O(d\eta^2 + L^{-3/2} \cdot \sigma^2 \cdot \eta)$$

Zou and Gu, On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients, ICML 2021

Application to Different Stochastic Gradient Estimators

- Stochastic gradients
  - Mini-batch stochastic gradient (SG)
  - Stochastic variance reduced gradient (SVRG) [Johnson and Zhang, 2013]
  - Stochastic averaged gradient (SAGA) [Defazio et. al., 2013]
  - Control variate gradient (CVG) [Baker et. al., 2018]
- Warm start: the initial point $\mathbf{x}_0$ is found via SGD such that $\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 = O(d/\mu)$.
- Additional Assumptions
  - $f_i(\mathbf{x})$ is L/n-smooth
  - $L, \mu = O(n)$

Algorithm 2 Stochastic Gradient Estimators

input: Current point $\mathbf{q}_k$, index of the HMC proposal t, random sampled mini-batch $\mathcal{I}_k$

Mini-batch Stochastic gradient

$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k} \nabla f_i(\mathbf{q}_k)$

Stochastic variance reduced gradient

if $k + Kt \mod N = 0$ **then**
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \nabla f(\mathbf{q}_k)$, $\widetilde{\mathbf{q}} = \mathbf{q}_k$
else
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k} \left[\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widetilde{\mathbf{q}})\right] + f(\widetilde{\mathbf{q}})$
end if

Stochastic averaged gradient

if $k + Kt = 0$ **then**
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \nabla f(\mathbf{q}_k)$, $\mathbf{G} = \{\nabla f_i(\mathbf{q}_k)\}_{i=1,\ldots,n}$
else
  $\widetilde{\mathbf{g}}_k = \sum_{i=1}^{n} \mathbf{G}_i$
  $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \frac{n}{B}\sum_{i\in\mathcal{I}_k} \left[\nabla f_i(\mathbf{q}_k) - \mathbf{G}_i\right] + \widetilde{\mathbf{g}}_k,$
  Update $\mathbf{G}_i \leftarrow \nabla f_i(\mathbf{q}_k)$ for all $i \in \mathcal{I}_k$
end if

Control variate gradient

$\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k) = \nabla f(\widehat{\mathbf{q}}) + \frac{n}{B}\sum_{i\in\mathcal{I}_k} [\nabla f_i(\mathbf{q}_k) - \nabla f_i(\widehat{\mathbf{q}})]$

output: $\mathbf{g}(\mathbf{q}_k, \boldsymbol{\xi}_k)$

Variance of Different Stochastic Gradient Estimators

- **Mini-batch stochastic gradients**

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{q}_k,\xi_k)-\nabla f(\mathbf{q}_k)\|_2^2\right]=\mathbb{E}\left[\left\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}\nabla f_i(\mathbf{q}_k)-\sum_{i=1}^n\nabla f_i(\mathbf{q}_k)\right\|_2^2\right]$$

$$\leq\frac{n^2}{B}\boxed{\mathbb{E}\left[\left\|\nabla f_i(\mathbf{q}_k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(\mathbf{q}_k)\right\|_2^2\right]}=O\Big(\mathbb{E}[\|\nabla f_i(\mathbf{x}^*)\|_2^2]\Big)$$

which we assume to be bounded by $O(d)$

- **Stochastic variance-reduced gradients**

$$\mathbb{E}\left[\|\mathbf{g}(\mathbf{q}_k,\xi_k)-\nabla f(\mathbf{q}_k)\|_2^2\right]=\mathbb{E}\left[\left\|\frac{n}{B}\sum_{i\in\mathcal{I}_k}[\nabla f_i(\mathbf{q}_k)-\nabla f_i(\tilde{\mathbf{q}})]+\nabla f(\tilde{\mathbf{q}})-\nabla f(\mathbf{q}_k)\right\|_2^2\right]$$

$$\leq\frac{n^2}{B}\boxed{\mathbb{E}\left[\left\|\nabla f_i(\mathbf{q}_k)-\nabla f_i(\tilde{\mathbf{q}})\right\|_2^2\right]}$$

$\tilde{\mathbf{q}}=\mathbf{q}_u$ for some $u\in[k-N,k-1]$

$$\leq\frac{L^2}{B}\boxed{\mathbb{E}\left[\|\mathbf{q}_k-\tilde{\mathbf{q}}\|_2^2\right]}=O(N^2d\eta^2)$$

Convergence Rates of Stochastic Gradient HMC

Theorem [Zou and Gu, 2021] Suppose all assumptions are satisfied, set $K = 1/(4\sqrt{L}\eta)$, then,

$$\mathscr{W}_2^2\big(\mathbf{P}(\mathbf{x}_T), \pi\big) \le e^{-T/(32\kappa)} \cdot \mathbb{E}\big[\|\mathbf{x}_0 - \mathbf{x}^\pi\|_2^2\big] + O(d\eta^2 + L^{-3/2} \cdot \sigma^2 \cdot \eta)$$

- Mini-batch SG-HMC $\quad \sigma^2 = O(B^{-1}n^2 d)$
- SVRG-HMC $\quad \sigma^2 = O(B^{-1}L^2N^2 d\eta^2)$
- SAGA-HMC $\quad \sigma^2 = O(B^{-3}L^2n^2 d\eta^2)$
- CVG-HMC $\quad \sigma^2 = O\big(B^{-1}Ld\big)$

Zou and Gu, On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients, ICML 2021

Comparison of Gradient Complexities

- Number of stochastic gradient calculations such that $\mathscr{W}_2(\mathbf{P}(\mathbf{x}_T), \pi) \leq \epsilon/\sqrt{n}$, where $L, \mu = O(n)$.

Algorithm	Query Complexity	Type
SGLD [Dalalyan and Karagulyan, 2019]	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$	LD
SVRG/SAGA-LD [Zou et. al., 2018b]	$\tilde{O}\left(\frac{n}{\epsilon}\right)$	LD
SG-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$	HMC
SVRG/SAGA-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\right)$	HMC
CVG-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	HMC

Dalalyan and Karagulyan, User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Processes and their Applications, 2019.

Zou et. al., Subsampled stochastic variance-reduced gradient Langevin dynamics UAI 2018b

Underdamped Langevin MCMC with Stochastic Gradients

- SDE description

γ :Friction parameter, u : inverse mass

$$\mathrm{d}\mathbf{v}(t) = -\gamma\mathbf{v}(t)\mathrm{d}t - u\nabla f(\mathbf{x}(t))\mathrm{d}t + \sqrt{2\gamma u}\cdot \mathrm{d}\mathbf{B}(t) \qquad \mathrm{d}\mathbf{x}(t) = \mathbf{v}(t)\mathrm{d}t$$

- Partially solve the SDE [Cheng et. al., 2018]

$$\mathbf{v}(t) = e^{-\gamma t}\cdot\mathbf{v}(0) - u\int_0^t e^{-\gamma(t-s)}\nabla f(\mathbf{x}(s))\mathrm{d}s + \sqrt{2\gamma u}\cdot\int_0^t e^{-\gamma(t-s)}\mathrm{d}\mathbf{B}(s)$$

$$\mathbf{x}(t) = \mathbf{x}(0) + \frac{1-e^{-\gamma t}}{\gamma}\mathbf{v}(0) + \int_0^t u\int_0^r e^{-\gamma(r-s)}\nabla f(\mathbf{x}(s))\mathrm{d}s\mathrm{d}r + \sqrt{2\gamma u}\cdot\int_0^t\int_0^r e^{-\gamma(r-s)}\mathrm{d}\mathbf{B}(s)\mathrm{d}r$$

- Can be exactly calculated
- Cannot be exactly calculated via stochastic gradient

- Discrete update using stochastic gradient ($u = 1/L, \gamma = 2$)

$$\mathbf{v}_{k+1} = e^{-\gamma\eta}\cdot\mathbf{v}_k - u\int_0^\eta e^{-\gamma(\eta-s)}\mathbf{g}(\mathbf{x}_k,\xi_k)\mathrm{d}s + \sqrt{2\gamma u}\cdot\int_0^\eta e^{-\gamma(\eta-s)}\mathrm{d}\mathbf{B}(s)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{1-e^{-\gamma\eta}}{\gamma}\mathbf{v}_k + \int_0^\eta u\int_0^r e^{-\gamma(r-s)}\mathbf{g}(\mathbf{x}_k,\xi_k)\mathrm{d}s\mathrm{d}r + \sqrt{2\gamma u}\cdot\int_0^\eta\int_0^r e^{-\gamma(r-s)}\mathrm{d}\mathbf{B}(s)\mathrm{d}r$$

Cheng et. al., Underdamped Langevin MCMC: A non-asymptotic analysis, COLT 2018

Convergence Analysis Framework

- Define 3 sequences:

$$(\mathcal{S}_\eta \mathbf{x}_k, \mathcal{S}_\eta \mathbf{v}_k) = (\mathbf{x}_{k+1}, \mathbf{v}_{k+1})$$

ULD with stochastic gradient

$$(\mathcal{G}_\eta \mathbf{x}_k, \mathcal{G}_\eta \mathbf{v}_k) = (\mathbb{E}[\mathbf{x}_{k+1} \,|\, \mathbf{x}_k, \mathbf{v}_k], \mathbb{E}[\mathbf{v}_{k+1} \,|\, \mathbf{x}_k, \mathbf{v}_k])$$

gradient ULD update

$$(\mathcal{L}_\eta \mathbf{x}_k, \mathcal{L}_\eta \mathbf{v}_k) = \left(\mathbf{x}_k + \int_0^\eta \mathbf{v}(s)\mathrm{d}s, \mathbf{v}_k - \int_0^\eta [-\gamma \mathbf{v}(s) - u\nabla f(\mathbf{x}(s))]\mathrm{d}s + \sqrt{2\gamma u}\int_0^\eta \mathrm{d}\mathbf{B}(s)\right)$$

Update via exact SDE solution

- Sampling error: we want to characterize the difference between $\mathcal{S}_\eta^T \mathbf{x}_0$ and x^π.

Sampling Error Decomposition

- Let $\mathbf{z}_k = \begin{pmatrix} \mathbf{x}_k \\ \mathbf{x}_k + \mathbf{v}_k \end{pmatrix} = \mathcal{S}_\eta^k \begin{pmatrix} \mathbf{x}_0 \\ \mathbf{x}_0 + \mathbf{v}_0 \end{pmatrix}$ and $\mathbf{z}^\pi = \begin{pmatrix} \mathbf{x}^\pi \\ \mathbf{x}^\pi + \mathbf{v}^\pi \end{pmatrix}$

$$\mathbb{E}\left[\|\mathbf{z}_k - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right] = \mathbb{E}\left[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 + \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right]$$
$$= \mathbb{E}\left[\|\mathcal{S}_\eta^k \mathbf{z}_0 - \mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0\|_2^2\right] + \mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right]$$

One-step statistical error between $\mathcal{S}_\eta$ and $\mathcal{G}_\eta$: $= O(L^{-2} \cdot \sigma^2 \cdot \eta^2)$

$$\mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right] = \mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 + \mathcal{L}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right]$$
$$= (1+\alpha)\mathbb{E}\left[\|\mathcal{L}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta^k \mathbf{z}^\pi\|_2^2\right]$$
$$+ (1 + 1/\alpha)\mathbb{E}\left[\|\mathcal{G}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0 - \mathcal{L}_\eta \mathcal{S}_\eta^{k-1} \mathbf{z}_0\|_2^2\right]$$

One-step discretization error between $\mathcal{G}_\eta$ and $\mathcal{L}_\eta$: $= O(\mu^{-1} d \cdot \eta^4)$

Contraction Property

- $\mathscr{L}_\eta$ has a good contraction property for any two points $\mathbf{z}$ and $\mathbf{z}'$ [Cheng et. al., 2018]

$$\mathbb{E}\left[\|\mathscr{L}_\eta \mathbf{z} - \mathscr{L}_\eta \mathbf{z}'\|_2^2\right] \leq e^{-\eta/\kappa} \cdot \|\mathbf{z} - \mathbf{z}'\|_2^2$$

- Error decomposition (set $\alpha = \eta/(2\kappa)$)

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{z}_k - \mathscr{L}_\eta^k \mathbf{z}^\pi\|_2^2\right] &\leq e^{-\eta/\kappa} \cdot (1+\alpha) \cdot \mathbb{E}\left[\|\mathbf{z}_{k-1} - \mathscr{L}_\eta^{k-1} \mathbf{z}^\pi\|_2^2\right] \\ &\quad + (1 + 1/\alpha) \cdot O(d \cdot \eta^4) + O(L^{-2} \cdot \sigma^2 \cdot \eta^2) \\ &\leq e^{-k\eta/(2\kappa)} \cdot \mathbb{E}\left[\|\mathbf{z}_0 - \mathbf{z}^\pi\|_2^2\right] + O(\mu^{-1} d \cdot \eta^2) + O(L^{-2} \cdot \sigma^2 \cdot \eta)\end{aligned}$$

Cheng et. al., Underdamped Langevin MCMC: A non-asymptotic analysis, COLT 2018

Convergence Rates of Stochastic Gradient ULD

Theorem [Zou et. al., 2018a, Chatterji et. al., 2018] Suppose all assumptions are satisfied, then,

$$\mathcal{W}_2^2\big(\mathbf{P}(\mathbf{x}_T), \pi\big) \le \big(1 - \eta/(2\kappa)\big)^T \cdot \mathbb{E}\big[\|\mathbf{x}_0 - \hat{\mathbf{x}}^\pi\|_2^2\big] + O(\mu^{-1} d \cdot \eta^2 + L^{-2} \cdot \sigma^2 \cdot \eta)$$

- Mini-batch SG-ULD $\quad \sigma^2 = O(B^{-1} n^2 d)$
- SVRG-ULD $\quad \sigma^2 = O(B^{-1} L^2 N^2 d \eta^2)$
- SAGA-ULD $\quad \sigma^2 = O(B^{-3} L^2 n^2 d \eta^2)$
- CVG-ULD $\quad \sigma^2 = O\big(B^{-1} L d\big)$

Zou et. al., Stochastic variance-reduced Hamilton Monte Carlo methods, ICML 2018

Chatterji et. al., On the Theory of Variance Reduction for Stochastic Gradient Monte Carlo, ICML 2018

Comparison of Gradient Complexities

- Number of stochastic gradient calculations such that $\mathscr{W}_2(\mathbf{P}(\mathbf{x}_T), \pi) \leq \epsilon/\sqrt{n}$, where $L, \mu = O(n)$.

Algorithm	Query Complexity	Type
SGLD [Dalalyan and Karagulyan, 2019]	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$	LD
SVRG/SAGA-LD [Zou et. al., 2018b]	$\tilde{O}\left(\frac{n}{\epsilon}\right)$	LD
SG-ULD [Chatterji et. al., 2018]	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$	ULD
SVRG/SAGA-ULD [Zou et. al., 2018a]	$\tilde{O}\left(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\right)$	ULD
CVG-ULD [Chatterji et. al., 2018]	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	ULD
SG-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{n}{\epsilon^2}\right)$	HMC
SVRG/SAGA-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{n^{2/3}}{\epsilon^{2/3}} + \frac{1}{\epsilon}\right)$	HMC
CVG-HMC [Zou and Gu, 2021]	$\tilde{O}\left(\frac{1}{\epsilon^2}\right)$	HMC

Summary

- We provided a unified analysis for HMC and ULD with stochastic gradients.
- The analysis is based on three sequences of Markov chains:
  - Markov chain of the stochastic gradient MCMC
  - Markov chain of the conditional expected stochastic gradient MCMC
  - Markov chain of the idealized HMC/ULD
- The analyses are different since HMC and ULD has different contraction property:
  - ULD has contraction property for any two points (so can be used in every iteration)
  - HMC has contraction property for any two points with the same velocity (so can only be used in every K iterations)

What's next?

- If the target distribution is not log-concave, the contraction property does not hold. Then how to control the approximation error of numerical solvers?
  - Show that the target distribution satisfies log-sobolev or Poincare inequality, which can give a weaker version of the contraction [Raginsky et. al., 2017, Vempala and Wibisono, 2019, Xu et al., 2018, Ma et. al., 2019, Zou et. al., 2021].
- Metropolis-Hasting step is skipped when using stochastic gradients, is it possible to approximately estimate this accept/reject probability to improve the sampling accuracy?
  - Develop an (nearly) unbiased estimator of the MH probability using the randomly sampled mini-batch data [Lee et. al., 2021]

Reference I

- Mengersen, K. L., Tweedie, R. L., et al. Rates of convergence of the Hastings and Metropolis algorithms. The annals of Statistics, 24(1):101–121, 1996.
- Lovasz, L. and Simonovits, M. The mixing rate of Markov Chains, an isoperimetric inequality, and computing the volume. In Proceedings [1990] 31st annual symposium on foundations of computer science, pp. 346–354. IEEE, 1990.
- Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid Monte Carlo. Physics letters B, 195(2):216–222, 1987.
- Roberts, G. O. and Tweedie, R. L. Exponential convergence of Langevin distributions and their discrete approximations. Bernoulli, pp. 341–363, 1996.
- Chen, Z. and Vempala, S. S. Optimal convergence rate of Hamiltonian Monte Carlo for strongly logconcave distributions. Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, 2019.
- Cheng, X., Chatterji, N. S., Abbasi-Yadkori, Y., Bartlett, P. L., and Jordan, M. I. Underdamped Langevin MCMC: A non-asymptotic analysis. In Conference on learning theory, pp. 300-323. PMLR, 2018
- Chatterji, N. S., Flammarion, N., Ma, Y.-A., Bartlett, P. L., and Jordan, M. I. On the theory of variance reduction for stochastic gradient Monte Carlo. In Proceedings of the 35th International Conference on Machine Learning, pp. 6028–6037, 2018
- Zou, D., Xu, P., and Gu, Q. Stochastic variance-reduced Hamilton Monte Carlo methods. In Proceedings of the 35th International Conference on Machine Learning, pp. 6028–6037, 2018.
- Zou, D. and Gu, Q. On the Convergence of Hamiltonian Monte Carlo with Stochastic Gradients. In Proceedings of the 38th International Conference on Machine Learning, pp. 6028–6037, 2021.

Reference II

- Johnson, R. and Zhang, T. Accelerating stochastic gradient descent using predictive variance reduction. In Advances in Neural Information Processing Systems, pp. 315–323, 2013.
- Defazio, A., Bach, F., and Lacoste-Julien, S. Saga: A fast incremental gradient method with support for nonstrongly convex composite objectives. In Advances in Neural Information Processing Systems, pp. 1646–1654, 2014.
- Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Conference on Learning Theory, pp. 1674–1703, 2017.
- Dalalyan, A. S. and Karagulyan, A. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. Stochastic Processes and their Applications, 129(12): 5278–5311, 2019.
- Vempala, S. and Wibisono, A. Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices." Advances in neural information processing systems 32 (2019): 8094-8106.
- Zou, D., Xu, P., and Gu, Q. Subsampled stochastic variance reduced gradient Langevin dynamics. In Proceedings of International Conference on Uncertainty in Artificial Intelligence, 2018.
- Ma, Y.-A., Chen, Y., Jin, C., Flammarion, N., and Jordan, M. I. Sampling can be faster than optimization. Proceedings of the National Academy of Sciences, 116(42), 20881-20885.
- Lee, Y.T., Shen, R. and Tian, K. Structured logconcave sampling with a restricted gaussian oracle. In Conference on Learning Theory, pp. 2993-3050, 2021.
- Zou, D., Xu, P., and Gu, Q. Faster Convergence of Stochastic Gradient Langevin Dynamics for Non-Log-Concave Sampling. In Proceedings of International Conference on Uncertainty in Artificial Intelligence, 2021.