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 $\pi(d\mathbf{x})$.

- In many cases, the target distribution is represented by $\pi \propto e^{-f(\mathbf{x})}$, where the negative log-density function $f(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}$ 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 $\pi$ is typically the posterior given i.i.d. observations $\{z_i\}_{i=1,\ldots,n}$.

$$\pi = p(x \mid z_1, \ldots, z_n) \propto \frac{p(z_1, \ldots, z_n \mid x) \cdot p(x)}{\prod_{i=1}^{n} p(z_i \mid x)}$$

Then $\pi$ can be rewritten as

$$\pi \propto e^{-f(x)} = e^{-\sum_{i=1}^{n} f_i(x)}$$

where

$$f_i(x) = -\log(p(z_i \mid x)) - n^{-1} \cdot \log(p(x))$$
Markov Chain Monte Carlo methods

- MCMC method
  - For $t = 1, \ldots, T$
    - **Proposal:** $x_{t+1} = x_t + g_f(x_t)$
      - A random vector depending on $f$ and $x_t$
    - **Reject:** $x_{t+1} = x_t$ with probability $1 - \alpha_f(x_t, x_{t+1})$
      - 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  \( H(x, p) = f(x) + \|p\|_2^2/2 \)

\[
\begin{align*}
\frac{dx(t)}{dt} &= \frac{\partial H(x(t), p(t))}{\partial p} = p(t) \\
\frac{dp(t)}{dt} &= - \frac{\partial H(x(t), p(t))}{\partial x} = -\nabla f(x(t))
\end{align*}
\]

- (Idealized) Hamiltonian Monte Carlo Method

  - \( x_{t+1} = x_t + \int_{\tau=0}^{\tau_0} p(\tau) \, d\tau \), where \( x(0) = x_t, \ p(0) \sim N(0, I) \)

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

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

- **SDE description**  
  \[ \begin{align*}  
dv(t) &= -\gamma v(t)dt - u \nabla f(x(t)) dt + \sqrt{2\gamma u} \cdot dB(t) \\
    dx(t) &= v(t)dt  
\end{align*} \]

- (Idealized) Underdamped Langevin MCMC Method
  
  \[ \begin{align*}  
x_{t+1} &= x_t + \int_{\tau=0}^{\eta} v(\tau) d\tau, \\
v_{t+1} &= v_t + \int_{\tau=0}^{\eta} \left[ \gamma v(\tau) + u \nabla f(x(\tau)) \right] d\tau + \sqrt{2\gamma u} \eta \cdot \xi_t
\end{align*} \]

  where \( v(0) = v_t, x(0) = x_t, \xi_t \sim N(0, I) \)

  **Key property:** When \( t \to \infty, \ (x_t, v_t) \sim \pi \propto e^{-f(x) - \|v\|^2/2} \)

MCMC with Stochastic Gradients

- Both HMC and underdamped LMC involve the calculation of the gradient \( \nabla f(x) \), 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 \( x_0, \eta, T, K \)
  - For \( t = 0, \ldots, T \)
    - Let \( p_0 \sim \mathcal{N}(0,I) \)
    - Let \( q_0 = x_t \)
    - For \( k = 0, \ldots, K - 1 \)
      - \( p_{k+1/2} = p_k - \frac{\eta}{2} g(q_k, \xi_k) \)
      - \( q_{k+1} = q_k + \eta p_{k+1/2} \)
      - \( p_{k+1} = p_k - \frac{\eta}{2} g(q_{k+1}, \xi_{k+1/2}) \)
  - Proposal: Numerically solving Hamilton’s equation via stochastic gradients \( g(q_k, \xi_k) \)
  - Leapfrog numerical integrator

- Output \( x_T \)

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

Approximation Error of the Numerical ODE Solver (Inner Loop)

- Define 3 sequences \((q_0 = x_t)\):
  
  \[
  (\mathcal{S}_\eta q_k, \mathcal{S}_\eta p_k) = (q_{k+1}, p_{k+1})
  
  (\mathcal{G}_\eta q_k, \mathcal{G}_\eta p_k) = (\mathbb{E}[q_{k+1} | p_k, q_k], \mathbb{E}[p_{k+1} | p_k, q_k])
  
  (\mathcal{H}_\eta q_k, \mathcal{H}_\eta p_k) = \left( q_k + \int_0^\eta p(t)dt, p_k - \int_0^\eta \nabla f(q(t))dt \right)
  
  HMC with stochastic gradient
  Conditionally expected stochastic gradient HMC update
  Update via exact ODE solution

- Approximation error: we want to characterize the difference between \(\mathcal{S}_\eta^K q_0\) and \(\mathcal{H}_\eta^K q_0\).
Decomposition of the Approximation Error (Inner Loop)

Define $z_k = \begin{pmatrix} q_k \\ L^{-1/2} p_k \end{pmatrix} = S^k_{\eta} \begin{pmatrix} q_0 \\ L^{-1/2} p_0 \end{pmatrix} = S^k_{\eta} z_0$, then

$E_k := \mathbb{E} \left[ \| S^k_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right] = \mathbb{E} \left[ \| S^k_{\eta} z_0 - G_{\eta} S^{k-1}_{\eta} z_0 + G_{\eta} S^{k-1}_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right]$

$= \mathbb{E} \left[ \| S^k_{\eta} z_0 - G_{\eta} S^{k-1}_{\eta} z_0 \|_2^2 \right] + \mathbb{E} \left[ \| G_{\eta} S^{k-1}_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right]$

One-step statistical error between $S_{\eta}$ and $G_{\eta}$: $= O(L^{-1} \cdot \sigma^2 \cdot \eta^2)$

$\mathbb{E} \left[ \| G_{\eta} S^{k-1}_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right] = \mathbb{E} \left[ \| G_{\eta} S^{k-1}_{\eta} z_0 - H_{\eta} S^{k-1}_{\eta} z_0 + H_{\eta} S^{k-1}_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right]$

$\leq (1 + \alpha) \cdot \mathbb{E} \left[ \| H_{\eta} S^{k-1}_{\eta} z_0 - H^k_{\eta} z_0 \|_2^2 \right]$

$+ (1 + 1/\alpha) \cdot \mathbb{E} \left[ \| G_{\eta} S^{k-1}_{\eta} z_0 - H_{\eta} S^{k-1}_{\eta} z_0 \|_2^2 \right]$

One-step “discretization error” between $G_{\eta}$ and $H_{\eta}$: $= O(L d \cdot \eta^4)$
Decomposition of the Approximation Error (Inner Loop)

- **Bound on** \( \mathbb{E} \left[ \| \mathcal{H}_\eta \mathcal{S}_\eta^{k-1} z_0 - \mathcal{H}_\eta^k z_0 \|_2^2 \right] \)

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

\[
\mathbb{E} \left[ \| \mathcal{H}_\eta \mathcal{S}_\eta^{k-1} z_0 - \mathcal{H}_\eta^k z_0 \|_2^2 \right] \leq e^{2L^{1/2} \eta} \cdot \mathbb{E} \left[ \| \mathcal{S}_\eta^{k-1} z_0 - \mathcal{H}_\eta^{k-1} z_0 \|_2^2 \right] = e^{2L^{1/2} \eta} \cdot \mathcal{E}_{k-1}
\]
Upper Bound of the Approximation Error

- Putting things together

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

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

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

\[ \mathcal{E}_K = \mathbb{E}[\| \mathcal{S}_\eta^K q_0 - \mathcal{H}_\eta^K q_0 \|_2^2] \leq O(d\eta^2 + L^{-3/2} \cdot \sigma^2 \cdot \eta) \]
Convergence Analysis of Outer Loop

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

- Analysis framework:

<table>
<thead>
<tr>
<th>HMC with stochastic gradients</th>
<th>$x_0$</th>
<th>$S^K \eta x_0$</th>
<th>$S^{2K} \eta x_0$</th>
<th>...</th>
<th>$S^{TK} \eta x_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idealized HMC with stationary initialization</td>
<td>$x^\pi$</td>
<td>$H^K \eta x^\pi$</td>
<td>$H^{2K} \eta x^\pi$</td>
<td>...</td>
<td>$H^{TK} \eta x^\pi$</td>
</tr>
</tbody>
</table>

- Sampling error: we will characterize the difference between $S^{TK} \eta x_0$ and $H^{TK} \eta x^\pi$. 
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 $(q, p)$ and $(q', p)$, then for any \( 0 \leq t \leq 1/(2\sqrt{L}) \),

\[
\mathbb{E} \left[ \| \mathcal{H}_t q - \mathcal{H}_t q' \|_2^2 \right] \leq (1 - \mu^2) \| q - q' \|_2^2
\]

Strongly log-concave parameter

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

\[
\mathbb{E} \left[ \| \mathcal{S}_\eta^K q_0 - \mathcal{H}_\eta^K q_0' \|_2^2 \right] \leq (1 + \beta) \| \mathcal{H}_\eta^K q_0 - \mathcal{H}_\eta^K q_0' \|_2^2 + (1 + 1/\beta) \mathbb{E} \left[ \| \mathcal{S}_\eta^K q_0 - \mathcal{H}_\eta^K q_0 \|_2^2 \right]
\]

Contracting term

\[
\leq (1 - 1/(16\kappa)) \| q_0 - q_0' \|_2^2
\]

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,

\[
\mathcal{W}_2^2(P(x_T), \pi) \leq e^{-T/(32\kappa)} \cdot \mathbb{E}[\|x_0 - x^\pi\|_2^2] + O(d\eta^2 + L^{-3/2} \cdot \sigma^2 \cdot \eta)
\]
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 $x_0$ is found via SGD such that $\|x_0 - x^*\|^2 = O(d/\mu)$.

- Additional Assumptions
  - $f_i(x)$ is $L/n$-smooth
  - $L, \mu = O(n)$

### Algorithm 2 Stochastic Gradient Estimators

```
1: input: Current point $q_k$, index of the HMC proposal $t$, random sampled mini-batch $I_k$
2: Mini-batch Stochastic gradient
3: $g(q_k, \xi_k) = \frac{n}{B} \sum_{i \in I_k} \nabla f_i(q_k)$
4: Stochastic variance reduced gradient
5: if $k + Kt \mod N = 0$ then
6: $g(q_k, \xi_k) = \nabla f(q_k), \tilde{q} = q_k$
7: else
8: Stochastic averaged gradient
9: $g(q_k, \xi_k) = \frac{n}{B} \sum_{i \in I_k} [\nabla f_i(q_k) - \nabla f_i(\tilde{q})] + f(\tilde{q})$
10: end if
11: Control variate gradient
12: $g(q_k, \xi_k) = \nabla f(\tilde{q}) + \frac{n}{B} \sum_{i \in I_k} [\nabla f_i(q_k) - G_i] + \tilde{g}_k$
13: Update $G_i \leftarrow \nabla f_i(q_k)$ for all $i \in I_k$
14: end if
15: output: $g(q_k, \xi_k)$
```
Variance of Different Stochastic Gradient Estimators

**Mini-batch stochastic gradients**

\[
\mathbb{E} \left[ \| g(q_k, \xi_k) - \nabla f(q_k) \|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{B} \sum_{i \in \mathcal{I}_k} \nabla f_i(q_k) - \frac{1}{n} \sum_{i=1}^n \nabla f_i(q_k) \right\|_2^2 \right] \\
\leq \frac{n^2}{B} \mathbb{E} \left[ \left\| \nabla f_i(q_k) - \nabla f_i(\tilde{q}) \right\|_2^2 \right] + \nabla f(\tilde{q}) - \nabla f(q_k) \\
= O\left( \mathbb{E} \left[ \| \nabla f_i(x^*) \|_2^2 \right] \right) = O\left( \frac{1}{n} \sum_{i=1}^n \nabla f_i(q_k) \right)
\]

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

**Stochastic variance-reduced gradients**

\[
\mathbb{E} \left[ \| g(q_k, \xi_k) - \nabla f(q_k) \|_2^2 \right] = \mathbb{E} \left[ \left\| \frac{1}{B} \sum_{i \in \mathcal{I}_k} \nabla f_i(q_k) - \nabla f_i(\tilde{q}) \right\|_2^2 \right] + \nabla f(\tilde{q}) - \nabla f(q_k) \\
\leq \frac{n^2}{B} \mathbb{E} \left[ \left\| \nabla f_i(q_k) - \nabla f_i(\tilde{q}) \right\|_2^2 \right] + \nabla f(\tilde{q}) - \nabla f(q_k) \\
\leq \frac{L^2}{B} \mathbb{E} \left[ \| q_k - \tilde{q} \|_2^2 \right] = O(N^2 d \eta^2)
\]

\( \tilde{q} = q_u \) for some \( u \in [k - N, k - 1] \)
Convergence Rates of Stochastic Gradient HMC

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

$$\mathcal{W}_2^2(P(x_T), \pi) \leq e^{-T/(32\kappa)} \cdot \mathbb{E}\left[\|x_0 - x^\pi\|^2_2\right] + O(d\eta^2 + L^{-3/2} \cdot \sigma^2 \cdot \eta)$$

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

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 $\mathcal{W}_2(P(x_T), \pi) \leq \epsilon / \sqrt{n}$, where $L, \mu = O(n)$.

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

Zou et. al., Subsampled stochastic variance-reduced gradient Langevin dynamics UAI 2018b.
Underdamped Langevin MCMC with Stochastic Gradients

- **SDE description**

\[
dv(t) = -\gamma v(t)dt - u \nabla f(x(t))dt + \sqrt{2\gamma u} \cdot dB(t) \quad dx(t) = v(t)dt
\]

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

\[
v(t) = e^{-\gamma t} \cdot v(0) - u \int_0^t e^{-\gamma(t-s)} \nabla f(x(s))ds + \sqrt{2\gamma u} \cdot \int_0^t e^{-\gamma(t-s)}dB(s)
\]

\[
x(t) = x(0) + \frac{1 - e^{-\gamma t}}{\gamma} v(0) + \int_0^t \int_0^r e^{-\gamma(r-s)} \nabla f(x(s))dsdr + \sqrt{2\gamma u} \cdot \int_0^t \int_0^r e^{-\gamma(r-s)}dB(s)dr
\]

- **Can be exactly calculated**

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

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

\[
v_{k+1} = e^{-\gamma} \cdot v_k - u \int_0^\eta e^{-\gamma(\eta-s)} g(x_k, \xi_k)ds + \sqrt{2\gamma u} \cdot \int_0^\eta e^{-\gamma(\eta-s)}dB(s)
\]

\[
x_{k+1} = x_k + \frac{1 - e^{-\gamma}}{\gamma} v_k + \int_0^\eta \int_0^r e^{-\gamma(r-s)} g(x_k, \xi_k)dsdr + \sqrt{2\gamma u} \cdot \int_0^\eta \int_0^r e^{-\gamma(r-s)}dB(s)dr
\]

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

- Define 3 sequences:

\[
(S_\eta x_k, S_\eta v_k) = (x_{k+1}, v_{k+1})
\]

ULD with stochastic gradient

\[
(G_\eta x_k, G_\eta v_k) = (\mathbb{E}[x_{k+1} | x_k, v_k], \mathbb{E}[v_{k+1} | x_k, v_k])
\]

gradient ULD update

\[
(L_\eta x_k, L_\eta v_k) = \left(x_k + \int_0^\eta v(s)ds, v_k - \int_0^\eta [-\gamma v(s) - u \nabla f(x(s))]ds + \sqrt{2\gamma u} \int_0^\eta dB(s)\right)
\]

Update via exact SDE solution

- Sampling error: we want to characterize the difference between \(S_\eta^T x_0\) and \(x^\pi\).
Sampling Error Decomposition

Let $z_k = \left( \begin{array}{c} x_k \\ x_k + v_k \end{array} \right) = S^k_\eta \left( \begin{array}{c} x_0 \\ x_0 + v_0 \end{array} \right)$ and $z^\pi = \left( \begin{array}{c} x^\pi \\ x^\pi + v^\pi \end{array} \right)$

$$\mathbb{E}\left[ \|z_k - \mathcal{L}^k_\eta z^\pi\|_2^2 \right] = \mathbb{E}\left[ \|S^k_\eta z_0 - G_\eta S^k_\eta z_0 + G_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta z^\pi\|_2^2 \right]$$

$$= \mathbb{E}\left[ \|S^k_\eta z_0 - G_\eta S^k_\eta z_0\|_2^2 \right] + \mathbb{E}\left[ \|G_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta z^\pi\|_2^2 \right]$$

One-step statistical error between $S_\eta$ and $G_\eta$: $= O(L^{-2} \cdot \sigma^2 \cdot \eta^2)$

$$\mathbb{E}\left[ \|G_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta z^\pi\|_2^2 \right] = \mathbb{E}\left[ \|G_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta z_0 + \mathcal{L}^k_\eta z_0 - \mathcal{L}^k_\eta z^\pi\|_2^2 \right]$$

$$= (1 + \alpha)\mathbb{E}\left[ \|L^k_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta z^\pi\|_2^2 \right]$$

$$+ (1 + 1/\alpha)\mathbb{E}\left[ \|G_\eta S^k_\eta z_0 - \mathcal{L}^k_\eta S^k_\eta z_0\|_2^2 \right]$$

One-step discretization error between $G_\eta$ and $L_\eta$: $= O(\mu^{-1} d \cdot \eta^4)$
Contraction Property

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

$$\mathbb{E}\left[\|\mathcal{L}_\eta z - \mathcal{L}_\eta z'\|^2\right] \leq e^{-\eta/k} \cdot \|z - z'\|^2$$

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

$$\mathbb{E}\left[\|z_k - \mathcal{L}_\eta^k z^\pi\|^2\right] \leq e^{-\eta/k} \cdot (1 + \alpha) \cdot \mathbb{E}\left[\|z_{k-1} - \mathcal{L}_\eta^{k-1} z^\pi\|^2\right]$$

$$+ (1 + 1/\alpha) \cdot O(d \cdot \eta^4) + O(L^{-2} \cdot \sigma^2 \cdot \eta^2)$$

$$\leq e^{-k\eta/(2\kappa)} \cdot \mathbb{E}\left[\|z_0 - z^\pi\|^2\right] + O(\mu^{-1} d \cdot \eta^2) + O(L^{-2} \cdot \sigma^2 \cdot \eta)$$

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,

$$\mathcal{W}_2^2(P(x_T), \pi) \leq \left(1 - \eta/(2\kappa)\right)^T \cdot \mathbb{E}\left[\|x_0 - \hat{x}_\pi\|_2^2\right] + O(\mu^{-1}d \cdot \eta^2 + L^{-2} \cdot \sigma^2 \cdot \eta)$$

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

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 $\mathcal{W}_2(P(x_T), \pi) \leq \epsilon/\sqrt{n}$, where $L, \mu = O(n)$.

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