Stochastic Gradient Hamiltonian Monte Carlo Methods with Recursive Variance Reduction

Difan Zou and **Pan Xu** and **Quanquan Gu**
University of California, Los Angeles

NEURAL INFORMATION PROCESSING SYSTEMS

Problem Setup and Background

- **Problem** Sample from the target distribution $\pi \propto \exp\{-f(\mathbf{x})\}$
- **Hamiltonian Langevin dynamics** stochastic differential equation

$$d\boldsymbol{V}_t = -\gamma \boldsymbol{V}_t dt - u\nabla f(\boldsymbol{X}_t)dt + \sqrt{2\gamma u}d\boldsymbol{B}_t$$
$$d\boldsymbol{X}_t = \boldsymbol{V}_t dt$$

  where the parameters are
  - $\gamma > 0$ is called the friction parameter
  - $u > 0$ is the inverse mass.
  - $\boldsymbol{B}_t$ is a standard Brownian motion in $\mathbb{R}^d$
- **Asymptotic property** Under certain assumptions on $\nabla f(\mathbf{x})$, the Hamiltonian Langevin dynamics has an unique stationary distribution, i.e., $(\boldsymbol{X}_\infty, \boldsymbol{V}_\infty) \sim \pi_{x,v} \propto \exp\{-f(\mathbf{x}) - \|\mathbf{v}\|_2^2)/(2u)\}$

Sampling Algorithm

- **Density function**
  Target density $\pi \propto e^{-f(\mathbf{x})}$, with $f(\mathbf{x}) = 1/n\sum_{i=1}^n f_i(\mathbf{x})$
- **Stochastic Recursive Variance Reduced HMC**
  - **Update form**

$$\mathbf{v}_{k+1} = \mathbf{v}_k e^{-\gamma\eta} - u\gamma^{-1}(1-e^{-\gamma\eta})\mathbf{g}_k + \boldsymbol{\epsilon}_k^v$$
$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma^{-1}(1-e^{-\gamma\eta})\mathbf{v}_k + u\gamma^{-2}(\gamma\eta + e^{-\gamma\eta} - 1)\mathbf{g}_k + \boldsymbol{\epsilon}_k^x$$

    - $\mathbf{g}_k$ denotes the semi-stochastic gradient
    - ϵ_k^v and ϵ_k^x are Gaussian random vectors
  - **Semi-stochastic gradient**
    - $k \bmod L = 0$:
      $\mathbf{g}_k = 1/B_0\sum_{i\in\widetilde{\mathcal{B}}_k}\nabla f_i(\widetilde{\mathbf{x}}_k)$
    - $k \bmod L \neq 0$:
      $\mathbf{g}_k = 1/B\sum_{i\in\mathcal{B}_k}\left[\nabla f_i(\mathbf{x}_k) - \nabla f_i(\mathbf{x}_{k-1})\right] + \mathbf{g}_{k-1}$
  - **Random vectors**
    The covariance matrix of random vectors $\boldsymbol{\epsilon}_k^v$ and $\boldsymbol{\epsilon}_k^x$ satisfies

$$\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^v)^\top] = u(1-e^{-2\gamma\eta})\cdot\mathbf{I}$$
$$\mathbb{E}[\boldsymbol{\epsilon}_k^x(\boldsymbol{\epsilon}_k^x)^\top] = u\gamma^{-2}(2\gamma\eta + 4e^{-\gamma\eta} - e^{-2\gamma\eta} - 3)\cdot\mathbf{I}$$
$$\mathbb{E}[\boldsymbol{\epsilon}_k^v(\boldsymbol{\epsilon}_k^x)^\top] = u\gamma^{-1}(1 - 2e^{-\gamma\eta} + e^{-2\gamma\eta})\cdot\mathbf{I}$$

Convergence Results

- **Assumptions**
  - **Smoothness** Each component function $f_i(\cdot)$ satisfies

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \le M\|\mathbf{x}-\mathbf{y}\|_2, \quad \forall \mathbf{x},\mathbf{y}\in\mathbb{R}^d$$

  - **(m,b)-Dissipative** The sum function $f(\cdot)$ satisfies

$$\langle\nabla f(\mathbf{x}),\mathbf{x}\rangle \ge m\|\mathbf{x}\|_2^2 - b, \quad \forall\mathbf{x}\in\mathbb{R}^d$$

- **Convergence rate of SRVR-HMC**

$$\mathcal{W}_2(\mathbb{P}(\mathbf{x}_K),\pi) = O\left(\Gamma_1\left(\left(1+\frac{L}{B}\right)K\eta^3 + \frac{K\eta}{\gamma^2 B_0}\right)^{1/4} + \frac{e^{-\mu_* K\eta}}{\mu_*}\right)$$

  - L: epoch length of SRVR-HMC, B: mini-batch size, B_0: outer batch size and η: step size
  - $\Gamma_1 = \mathrm{poly}(d)$ and $\mu_* = e^{-O(d)}$ is the spectral gap of Hamiltonian Langevin dynamics
- **Convergence rate of SG-UL-MCMC**

$$\mathcal{W}_2(\mathbb{P}(\mathbf{x}_K),\pi) = O\left(\Gamma_1\left[2K\eta^3 + \frac{K\eta}{\gamma^2 B_0}\cdot\mathbb{1}(B_0<n)\right]^{1/4} + \frac{e^{-\mu_* K\eta}}{\mu_*}\right)$$

  - B_0: mini-batch size in each iteration
  - $\Gamma_1 = \mathrm{poly}(d)$ and $\mu_* = e^{-O(d)}$

  Remark: setting $B_0 = n$ implies the convergence rate of UL-MCMC

Comparison with the State-of-the-art

- **Gradient complexity**
  Number of stochastic gradient evaluations needed to achieve $\mathcal{W}_2(\mathbb{P}(\mathbf{x}_K),\pi) \le \epsilon$

Methods	Gradient Complexity
LMC	$\widetilde{O}(\epsilon^{-4}\lambda_*^{-5}n)$
SGLD	$\widetilde{O}(\epsilon^{-8}\lambda_*^{-9})$
SVRG-LD	$\widetilde{O}(n + \epsilon^{-2}\lambda_*^{-4}n^{3/4} + \epsilon^{-4}\lambda_*^{-4}n^{1/2})$
HMC	$\widetilde{O}(\epsilon^{-4}\mu_*^{-3}n)$
UL-MCMC	$\widetilde{O}(\epsilon^{-2}\mu_*^{-3/2}n)$
SGHMC	$\widetilde{O}(\epsilon^{-8}\mu_*^{-5})$
SG-UL-MCMC	$\widetilde{O}(\epsilon^{-6}\mu_*^{-5/2})$
SRVR-HMC	$\widetilde{O}((n + \epsilon^{-2}n^{1/2}\mu_*^{-3/2}) \wedge \epsilon^{-4}\mu_*^{-2})$

Sampling from Gaussian Mixture Distribution

- **Experiment setup**
  run all algorithms for 10^4 data passes and report
  1) density plot
  2) comparison in terms of mean square error

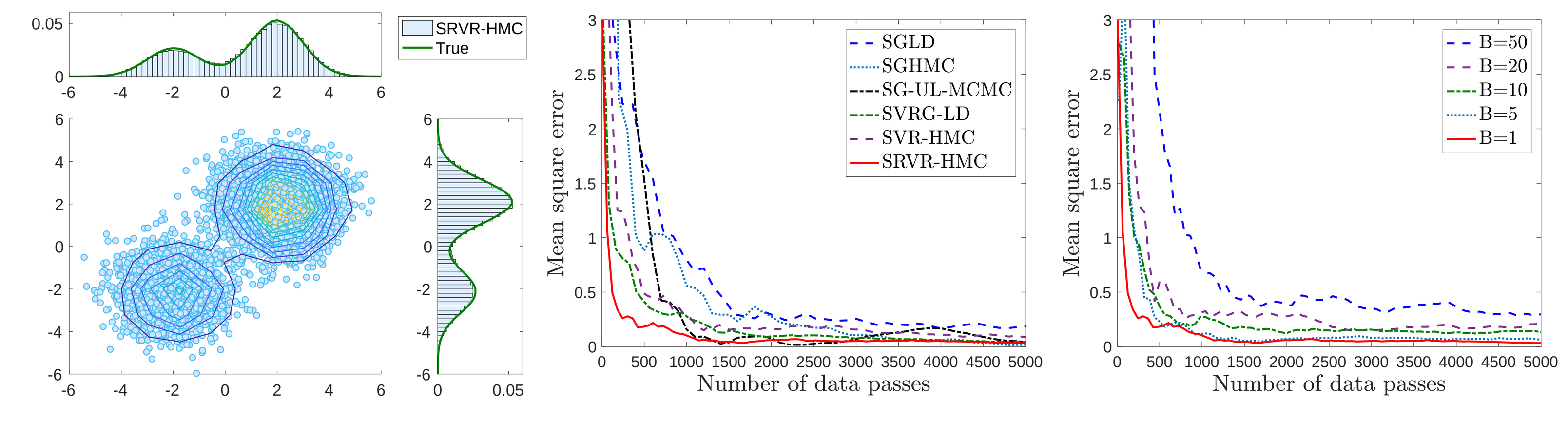
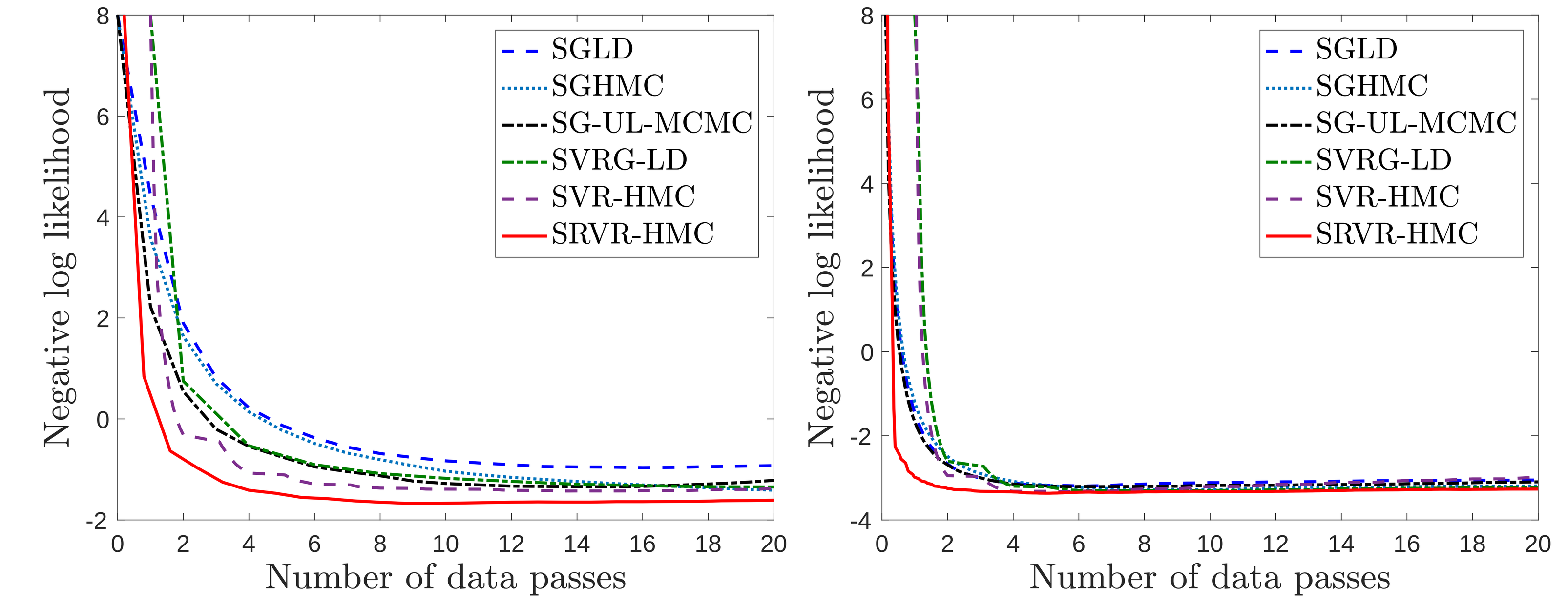


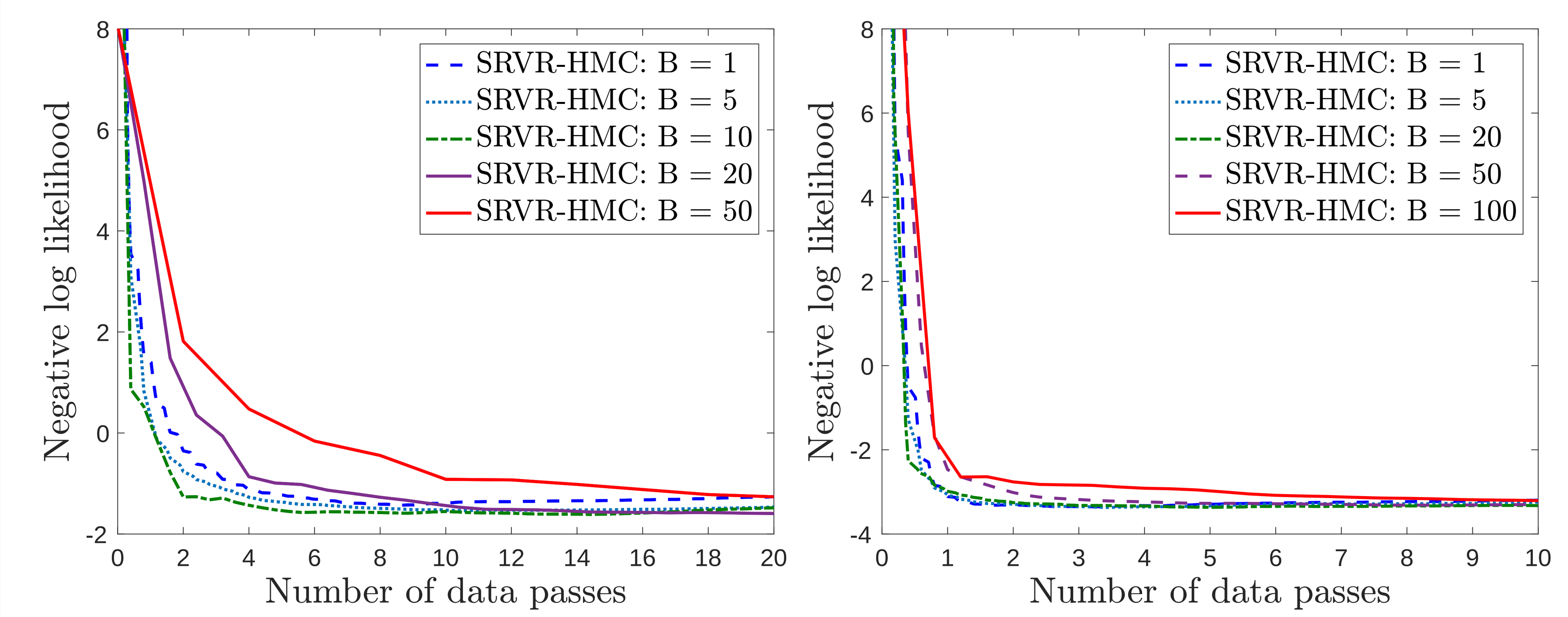
Figure: (a) 2D projection of kernel densities. (b) Comparison with baseline algorithms. (c) Convergence of SRVR-HMC with varying batch size B.

Independent Components Analysis

- **Experiment setup**
  - Training sample size $500/5000$. Test sample size 12730
  - Batch size $B_0 = n/5$, mini batch size $B = 10$ and epoch length $L = B_0/B$



(a) n=500, B_0=100 (b) n=5000, B_0=1000



(c) n=500, B_0=100 (d) n=5000, B_0=1000