Optimal Underdamped Langevin MCMC Method

Zhengmian Hu, Feihu Huang, Heng Huang

Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA.

Sampling by Underdamped Langevin MCMC

Problem: Sampling from $p(x) \propto exp(-\sum_{i=1}^{N} f_i(x))$.

Assumptions: $mI \preccurlyeq \nabla^2 f(x),\ \nabla^2 f_i(x) \preccurlyeq \frac{L}{N} I$.

(ULD) $dX_t = V_t dt,\ dV_t = -\nabla f(X_t)dt - \gamma V_t dt + \sqrt{2\gamma} dB_t$.

ULD MCMC: Markov chain by discretization of ULD.

Oracles: Gradient oracle $\nabla f_i(x)$. Weighted Brownian oracle.
No function oracle.

Gradient Complexity

Table: Number of gradient evaluation of $\nabla f_i(x)$ needed to sample from m-strongly-log-concave distributions up to $\varepsilon\sqrt{d/m}$ accuracy in 2-Wasserstein distance

Algorithms	Gradient complexities
ULA (A. Dalalyan, 2017; Durmus, Moulines, et al., 2019)	$\widetilde{O}(N\varepsilon^{-2})$
LPM (A. S. Dalalyan, Riou-Durand, et al., 2020)	$\widetilde{O}(N\varepsilon^{-1})$
RMM (Shen and Lee, 2019)	$\widetilde{O}(N\varepsilon^{-\frac{2}{3}})$
ALUM (Ours)	$\widetilde{O}(N\varepsilon^{-\frac{2}{3}})$
SG-LPM (Cheng et al., 2018)	$\widetilde{O}(\varepsilon^{-2})$
SVRG-LPM (Zou, Xu, and Gu, 2018)	$\widetilde{O}(N + \varepsilon^{-1} + N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$
SVRG-ALUM (Ours)	$\widetilde{O}(N + N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$
SAGA-ALUM (Ours)	$\widetilde{O}(N + N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$

AcceLerated ULD-MCMC (ALUM)

$$\text{(ULD)} \qquad dX_t = V_t dt,\; dV_t = -\nabla f(X_t)dt - \gamma V_t dt + \sqrt{2\gamma}dB_t.$$

Estimation at time point h by only single gradient evaluation:

$$\begin{aligned}
X_h^{(o)} &= X_0 + \psi_1(h)V_0 - h\psi_1(h-ah)\nabla f(X_{ah}^{(e)}) + e_{x,[0,h]},\\
V_h^{(o)} &= \psi_0(h)V_0 - h\psi_0(h-ah)\nabla f(X_{ah}^{(e)}) + e_{v,[0,h]},\\
X_{ah}^{(e)} &= X_0 + \psi_1(ah)V_0 \quad \cancel{-\psi_2(ah)\nabla f(X_0)} \quad + e_{x,[0,ah]},
\end{aligned}$$

Similar to RMM (Shen and Lee, 2019) but save half gradient evaluations.

Surprisingly, dropping this term doesn't hinder the convergence too much.
The asymptotic iteration complexity has same d, ε dependence. In high precision regime (ε is small enough), the κ dependence is also the same.

Variance Reduced ALUM (VR-ALUM)

$$\widetilde{\nabla}_k^{\text{SVRG}} = \frac{N}{b}\sum_{i \in B_k}\left(\nabla f_i(x_k^{(e)\widetilde{\nabla}}) - \nabla f_i(\overline{x})\right) + \sum_{i=1}^N \nabla f_i(\overline{x}).$$

(Johnson and Zhang, 2013)

$$\widetilde{\nabla}_k^{\text{SAGA}} = \frac{N}{b}\sum_{i \in B_k}\left(\nabla f_i(x_k^{(e)\widetilde{\nabla}}) - \nabla f_i(\phi_k^i)\right) + \sum_{i=1}^N \nabla f_i(\phi_k^i).$$

(Defazio, Bach, and Lacoste-Julien, 2014)

Bounded MSE property - control the gradient error for different gradient estimations in a unified approach.

$$\mathbb{E}[\|\widetilde{\nabla}_{k+1} - \nabla f(x_{k+1}^{(e)})\|_2^2] \leq \Theta \max_{0 \leq i \leq k} Q_i,$$

$$\|\widetilde{\nabla}_0 - \nabla f(x_0^{(e)})\|_2^2 = 0,$$

$$Q_k = N \sum_{i=1}^N \|\nabla f_i(x_{k+1}^{(e)}) - \nabla f_i(x_k^{(e)})\|_2^2.$$

Upper Bounds

Table: Iteration complexities for full gradient ALUM.

Problem	Accuracy	Iteration complexity
Sampling	$\varepsilon\sqrt{d/m}$ in W_2	$\widetilde{O}(\max(\kappa/\varepsilon^{\frac{2}{3}}, \kappa^2))$
Approximating	ε in $\mathbb{L}_2$	$O(\max(T\kappa^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}}d^{\frac{1}{3}}, T\kappa))$

Table: Gradient complexity for SAGA-ALUM and SVRG-ALUM.

Problem	Accuracy	Gradient complexity
Sampling	$\varepsilon\sqrt{d/m}$ in W_2	$\widetilde{O}(N + (b\kappa + N^{\frac{2}{3}}\kappa^{\frac{4}{3}})(1+\varepsilon^{-\frac{2}{3}}) + b\kappa^2$
Approximating	ε in $\mathbb{L}_2$	$O(N + T(\kappa b + \kappa^{\frac{1}{3}}N^{\frac{2}{3}}) + T\kappa^{\frac{2}{3}}d^{\frac{1}{3}}\varepsilon^{-\frac{2}{3}}(b + N^{\frac{2}{3}}))$

Corollary

When $b \leq O(N^{\frac{2}{3}})$, the gradient complexity of SAGA-ALUM and SVRG-ALUM for sampling problem is $\widetilde{O}(N + N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$ and their gradient complexity for ULD approximation problem is $O(N + d^{\frac{1}{3}}N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$.

Lower Bounds for Approximation Error

Problem class: $\mathcal{U}$ are all strongly convex and uniformly smooth functions f_i such that mean of $\frac{1}{Z}exp(-\sum_{i=1}^{N} f_i(x))$ is not too far from origin.
Single component gradient oracle: $\nabla f_i(x)$.
Weighted Brownian oracle: $\int_0^T e^{\theta s} dB_s(\omega)$.
Ground truth: $X_T(\omega, U)$.
All possible randomized algorithms with n evaluations of oracles: $\mathcal{A}_n$.
Worst case approximation error:

$$e^2_{\mathcal{A},\mathcal{U}} := \inf_{A\in\mathcal{A}} \sup_{U\in\mathcal{U}} \mathbb{E}_{\omega\in\mathbb{P}}\mathbb{E}_{\widetilde{\omega}\in\widetilde{\mathbb{P}}}\|X_T(\omega, U) - A(\omega, \widetilde{\omega}, U)\|_2^2$$

Lower Bounds for Approximation Error

Theorem

When $n < N$ which means that gradient evaluation number is less than components number, we have $e^2_{\mathcal{A}_n,\mathcal{U}} \geq dC_1$, where C_1 is positive and independent of d, N, and n.

Theorem

When gradient evaluation number n is multiple of N, we have $e^2_{\mathcal{A}_n,\mathcal{U}} \geq dC_2\frac{N^2}{n^3}$, where C_2 is positive and independent of d, N, and n.

Corollary

For small enough target accuracy ε such that $\varepsilon^2 < dC_1$, in order to achieve $e_{\mathcal{A}_n,\mathcal{U}} \leq \varepsilon$, the minimum number of single component gradient oracle evaluations is $\Omega(N + d^{\frac{1}{3}}N^{\frac{2}{3}}\varepsilon^{-\frac{2}{3}})$.

This lower bound matches the upper bound in the dependence of d, components number N, and approximation accuracy ε.

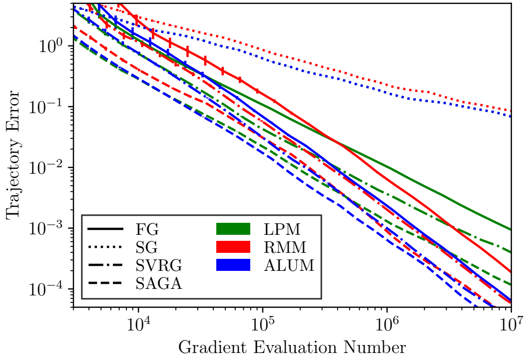
In what sense VR-ALUM is optimal (or not)?

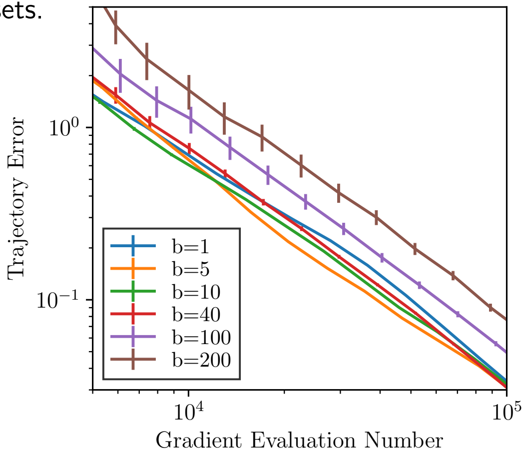
Optimal for approximating problem in the sense that any ULD MCMC algorithm with better dependence on dimension d, components number N, approximation accuracy ε in gradient complexity doesn't exist.

Not necessarily optimal in sampling error, κ dependence, or when other assumptions and oracles exist.

Experiments

Bayesian logistic regression on LIBSVM datasets.





VR-ALUM constantly outperforms other discretizations of ULD.

Approximating efficiency is not sensitive to batch size when batch is relatively small.

Thank you