Constrained Langevin Algorithms with L-mixing External Random Variables

Yuping Zheng & Andrew Lamperski

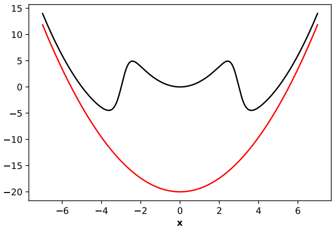
Electrical and Computer Engineering Department
University of Minnesota

Oct 18, 2022

Stochastic Optimization Problem

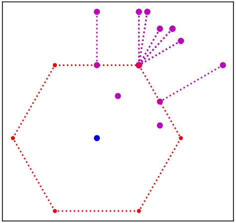
$$\min_{x \in \mathcal{K}} \ \bar{f}(x) = \mathbb{E}_z[f(x, \mathbf{z})]$$

Averaged Loss $\bar{f}(\mathbf{x})$



Strongly convex outside a ball

Constraint $\mathcal{K}$



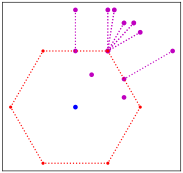
Polyhedral, not necessarily bounded

Gradient descent + Additive noise

Next State — Current State — Correlated External Data — IID $\mathcal{N}(0, I)$ Noise

$$\mathbf{x}_{k+1} = \Pi_{\mathcal{K}}\left(\mathbf{x}_k - \eta\nabla_x f(\mathbf{x}_k, \mathbf{z}_k) + \sqrt{\frac{2\eta}{\beta}}\hat{\mathbf{w}}_k\right)$$

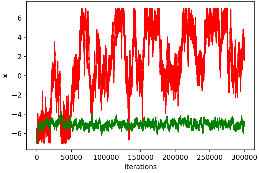
Convex Projection — Step Size — Loss Function — Inverse Temperature



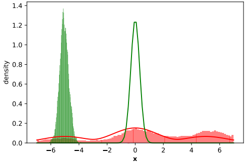
Motivation

- Why Langevin algorithms?
  - Potential choice for adaptive control, deep neural networks, reinforcement learning, time series analysis, image processing and so on
  - Impossible to find an algorithm that efficiently solves all the non-convex optimization problems
  - Properly-scaled additive noise assists to escape from local minima and saddles
- Why constraint?
  Polyhedral constraint is very common in applications with box and simplex constraints

Trajectories



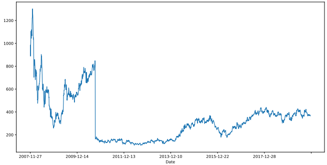
Histograms



Why L-mixing Processes?

- Minimizing a loss with external random variables is common
- External random variables are not necessarily IID
- The class of L-mixing processes was introduced in (Gerencsér, 1989) for system identification and time-series analysis
- The class of L-mixing processes gives a means to quantitatively measure how the dependencies between the random variables decay over time

Stock Market Data [1]



[1] https://www.kaggle.com/datasets/rohanrao/nifty50-stock-market-data

Related Work

Optimization

- Initial work: Gelfand and Mitter (1991); Borkar and Mitter (1999)
- Unconstrained: Raginsky, Rakhlin, and Telgarsky (2017); Chau et al. (2021)
- Constrained: Lamperski (2021); Sato et al. (2022)

Sampling

- Initial work: Roberts, Tweedie, et al. (1996)
- Strongly log-concave: Dalalyan (2017); Durmus, Moulines, et al. (2017)
- Only log-concave: Dalalyan, Karagulyan, and Riou-Durand (2019); Mou et al. (2019)
- Non log-concave: Majka, Mijatović, Szpruch, et al. (2020); Zou, Xu, and Gu (2021)
- Constrained: Bubeck, Eldan, and Lehec (2015); Bubeck, Eldan, and Lehec (2018); Hsieh et al. (2018); Ahn and Chewi (2020); Zhang et al. (2020)

Related Work

Learning

- Bayesian learning: Welling and Teh (2011)
- IID external variables: Raginsky, Rakhlin, and Telgarsky (2017); Lamperski (2021)
- Dependent external variables: Chau et al. (2021)
- Advanced Langevin algorithms: Girolami and Calderhead (2011); Ahn, Korattikara, and Welling (2012); Ma, Chen, and Fox (2015); Kim, Song, and Liang (2020)

Result & Comparison

Theorem 1

Assume that $\eta \le \min\left\{\frac{1}{4}, \frac{\mu}{4\ell^2}\right\}$, $\mathcal{K}$ is a polyhedron with 0 in its interior, $\mathbf{x}_0 \in \mathcal{K}$, and $\mathbb{E}[\|\mathbf{x}_0\|^2] \le \varsigma$. There are constants a, c_1, c_2, c_3, and c_4 such that the following bound holds for all integers $k \ge 4$:

$$W_1(\mathcal{L}(\mathbf{x}_k), \pi_{\beta\bar{f}}) \le (c_1 + c_2\sqrt{\varsigma})e^{-\eta a k} + (c_3 + c_4\sqrt{\varsigma})\sqrt{\eta \log(\eta^{-1})}.$$

In particular, if $\eta = \frac{\log T}{2aT}$, $T \ge 4$ and $T \ge e^{2a}$, then

$$W_1(\mathcal{L}(\mathbf{x}_T), \pi_{\beta\bar{f}}) \le \left(c_1 + c_2\sqrt{\varsigma} + \frac{c_3 + c_4\sqrt{\varsigma}}{(2a)^{1/2}}\right) T^{-1/2}\log T.$$

Gibbs distribution: $\pi_{\beta\bar{f}}(A) = \frac{\int_{A\cap\mathcal{K}} e^{-\beta\bar{f}(x)}dx}{\int_{\mathcal{K}} e^{-\beta\bar{f}(x)}dx}$

	Constraint	RV	Convergence Rate
Our work	noncompact	L-mixing	$O(T^{-1/2}\log T)$
Chau et al. (2021)	unconstrained	L-mixing	$O(T^{-1/2}(\log T)^{1/2})$
Lamperski (2021)	compact	IID	$O(T^{-1/4}(\log T)^{1/2})$