Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization

Lingjiong Zhu

Florida State University
Email: zhu@math.fsu.edu

AMS Fall Southeastern Sectional Meeting
20-21 November 2021
Joint with Xuefeng Gao (Chinese University of Hong Kong)
Mert Gürbüzbalaban (Rutgers University)

Population risk minimization

Consider the stochastic non-convex optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) := \mathbb{E}_{Z \sim \mathcal{D}}[f(x, Z)]. \tag{1}$$

- Z is a random variable whose probability distribution $\mathcal{D}$ is unknown, supported on some unknown set $\mathcal{Z}$.
- Functions $x \mapsto f(x, z)$ are continuous and can be non-convex.
- Having access to i.i.d samples $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$ where $Z_i \sim \mathcal{D}$, the goal is to generate an approximate minimizer X_k (possibly random) with small expected excess risk:

$$\mathbb{E}F(X_k) - F^*, \tag{2}$$

where $F^* = \min_x F(x)$ is the minimum value, and the expectation is taken with respect to both $\mathbf{Z}$ and X_k.

Empirical risk minimization

As $\mathcal{D}$ is unknown, it is natural to consider empirical risk minimization:

$$\min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x, z_i)\,, \tag{3}$$

based on the (deterministic) dataset $\mathbf{z} := (z_1, z_2, \ldots, z_n) \in \mathcal{Z}^n$ as a proxy to the problem (1) and minimize

$$\mathbb{E}F_{\mathbf{z}}(X_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \tag{4}$$

approximately, where the expectation is taken with respect to any randomness encountered during the algorithm to generate X_k.

Applications

- Such stochastic non-convex optimization problems arise in many applications including machine learning.
- One prominent example is the training of deep neural networks, where non-convex optimization witnesses empirical successes.
  - $F_{\mathbf{z}}(x) : \mathbb{R}^d \to \mathbb{R}$ denotes the loss function, and $f(x, z_i) = \ell(g(x, a_i), y_i)$ the loss contributed by an individual data point $z_i = (a_i, y_i)$, $i \in \{1, \ldots, n\}$, $x \in \mathbb{R}^d$ the collection of all the parameters of the neural network.
  - In regression and classification problems such as logistic regression and support vector machines, f is convex; whereas in deep learning f is typically non-convex (Vapnik (2013)).

Non-convex optimization

- Many algorithms have been proposed to solve the problem (1) and its finite-sum version (3).
- Among these, gradient descent, stochastic gradient and its variance-reduced or momentum-based variants come with guarantees for finding a local minimizer or a stationary point for non-convex problems.
- In some applications, convergence to a local minimum can be satisfactory (Ge et al. (2017), Du et al. (2017)).
- However in general, methods with global convergence guarantees are also desirable and preferable in many settings (Hazan et al. (2016), Şimşekli et al. (2018)).

Gradient Descent for Non-Convex Objective

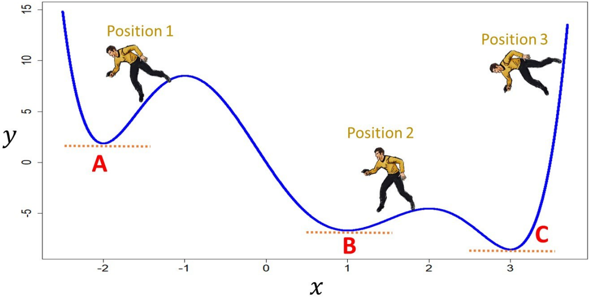

Figure: To solve $\min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) := \frac{1}{n}\sum_{i=1}^{n} f(x, z_i)$, the most common strategy is to use gradient descent: $X_{k+1} = X_k - \eta\nabla F_{\mathbf{z}}(X_k)$. For non-convex optimization problems, gradient descent algorithm can be stuck at a local minimum or stationary point.

Langevin based algorithms

- Stochastic gradient algorithms based on Langevin Monte Carlo are popular variants of stochastic gradient which admit asymptotic global convergence guarantees where a properly scaled Gaussian noise is added to the gradient updates.
- The properly scaled Gaussian noise term helps the Langevin algorithms to escape the local minima or stationary points.
- The algorithm will converge to a stationary distribution instead of a deterministic limit. The stationary distribution will concentrate around the global minimizer of $F_{\mathbf{z}}$.

SGLD and SGHMC

- Two popular Langevin-based algorithms that have demonstrated empirical success are
  - Stochastic gradient Langevin dynamics (SGLD) (Welling and Teh (2011))
  - Stochastic gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al. (2014), Chen et al. (2015), Neal (2010))
- Their variants have also been studied to improve their efficiency and accuracy (Ahn et al. (2012), Ma et al. (2015), Patterson and Teh (2013), Ding et al. (2014), Wibisono (2018)).

Overdamped Langevin SDE

- The first-order (a.k.a. overdamped) Langevin stochastic differential equation (SDE) is given by

$$dX(t) = -\nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\beta^{-1}}dB(t), \quad t \geq 0, \qquad (5)$$

where $\{B(t) : t \geq 0\}$ is the standard Brownian motion in $\mathbb{R}^d$.

- Under some assumptions on $F_{\mathbf{z}}$, the process X admits a unique stationary distribution $\pi_{\mathbf{z}}(dx) \propto \exp(-\beta F_{\mathbf{z}}(x))$, also known as the Gibbs measure.

- For β chosen large enough, it is easy to see that this Gibbs distribution $\pi_{\mathbf{z}}(dx)$ will concentrate around global minimizers of $F_{\mathbf{z}}$.

SGLD and Euler discretization of Overdamped SDE

- SGLD iterations consist of

$$X_{k+1} = X_k - \eta g_k + \sqrt{2\eta\beta^{-1}}\xi_k,$$

  - $\eta > 0$ is the stepsize parameter,
  - g_k is a conditionally unbiased estimate of the gradient of $\nabla F_{\mathbf{z}}(X_k)$,
  - β is the inverse temperature,
  - $(\xi_k)_{k=0}^{\infty}$ is a sequence of i.i.d standard Gaussian random vectors in $\mathbb{R}^d$.

- When the gradient variance is zero ($g_k = \nabla F_{\mathbf{z}}(X_k)$), SGLD dynamics corresponds to Euler discretization of overdamped Langevin SDE:

$$dX(t) = -\nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\beta^{-1}}dB(t).$$

Finite-time performance bounds for SGLD

- In a seminal work, Raginsky et al. (2017) [1] showed that SGLD iterates track the overdamped Langevin SDE closely and obtained finite-time performance bounds for SGLD.
- Related results also appear in Zhang et al. (2017) [2] and Xu et al. (2018) [3] .

[1] Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In: *Conference on Learning Theory*, pp 1674-1703.

[2] Zhang, Y. , Liang, Pl. and M. Charikar. (2017). A hitting time analysis of stochastic gradient Langevin dynamics. In: *Conference on Learning Theory*, pp 1674-1703.

[3] Xu, P., Chen, J., Zou, D. and Q. Gu. (2018). Global convergence of Langevin dynamics based algorithms for nonconvex optimization. In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Underdamped Langevin SDE

- The underdamped (second-order) Langevin SDE is given by:

$$dV(t) = -\gamma V(t)dt - \nabla F_{\mathbf{z}}(X(t))dt + \sqrt{2\gamma\beta^{-1}}dB(t), \tag{6}$$
$$dX(t) = V(t)dt, \tag{7}$$

- Under some assumptions on $F_{\mathbf{z}}$, the Markov process (X, V) is ergodic and have a unique stationary distribution

$$\pi_{\mathbf{z}}(dx, dv) = \frac{1}{\Gamma_{\mathbf{z}}} \exp\left(-\beta\left(\frac{1}{2}\|v\|^2 + F_{\mathbf{z}}(x)\right)\right) dxdv, \tag{8}$$

- Notice that the x-marginal distribution of $\pi_{\mathbf{z}}(dx, dv)$ is exactly the stationary distribution of the overdamped Langevin SDE.

SGHMC and Euler discretization of underdamped SDE

SGHMC algorithm is based on the discretization of underdamped (second-order) Langevin diffusion:

$$V_{k+1} = V_k - \eta[\gamma V_k + g(X_k, U_{\mathbf{z},k})] + \sqrt{2\gamma\beta^{-1}\eta}\xi_k, \tag{9}$$

$$X_{k+1} = X_k + \eta V_k. \tag{10}$$

- $(\xi_k)_{k=0}^{\infty}$ is a sequence of i.i.d standard Gaussian random vectors in $\mathbb{R}^d$,
- $\{U_{\mathbf{z},k} : k = 0, 1, \ldots\}$ is a sequence of i.i.d random elements such that $\mathbb{E}g(x, U_{\mathbf{z},k}) = \nabla F_{\mathbf{z}}(x)$ for any $x \in \mathbb{R}^d$.
- There is an alternative discretization (we call it SGHMC2) introduced by Cheng et al. (2017) with better diffusion approximation error.

Motivation

- In the optimization literature, it is well known that gradient descent with momentum, e.g. Nesterov's accelerated gradient descent can outperform gradient descent.
- Recent results of Eberle et al. (2019) [4] showed that underdamped SDE can converge to its stationary distribution faster than the overdamped SDE (in the 2-Wasserstein metric) under some assumptions where $F_{\mathbf{z}}$ can be non-convex.
- This raises the natural question whether the discretized underdamped dynamics (SGHMC), can lead to better guarantees than the SGLD method.

[4] Eberle, A., Guillin, A. and R. Zimmer (2019). Couplings and quantitative contraction rates for Langevin dynamics. *Annals of Probability*. 47:1982-2010.

Contributions

- We give first-time finite-time guarantees for SGHMC to find approximate minimizers of both empirical and population risks with explicit constants [5].

- We also show that on a class of non-convex problems, SGHMC can converge faster than SGLD by a square root factor.
  - Momentum-based acceleration is achievable for some classes of non-convex problems, as empirically observed in practice.
  - Bridge a gap between the theory and the practice for the use of SGHMC algorithms in stochastic non-convex optimization.

[5] Gao, X., Gürbüzbalaban, M. and Zhu, L. (2021+). Global Convergence of Stochastic Gradient Hamiltonian Monte Carlo for Non-Convex Stochastic Optimization: Non-Asymptotic Performance Bounds and Momentum-Based Acceleration. To appear in *Operations Research*.

Assumptions

(*i*) The function f is continuously differentiable, takes non-negative real values, and there exist constants $A_0, B \geq 0$ so that for any $z \in \mathcal{Z}$.

$$|f(0,z)| \leq A_0, \qquad \|\nabla f(0,z)\| \leq B.$$

(*ii*) For each $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is M-smooth:

$$\|\nabla f(w,z) - \nabla f(v,z)\| \leq M\|w - v\|.$$

(*iii*) For each $z \in \mathcal{Z}$, the function $f(\cdot, z)$ is (m,b)-dissipative:

$$\langle x, \nabla f(x,z)\rangle \geq m\|x\|^2 - b\,.$$

(*iv*) There exists a constant $\delta \in [0,1)$ such that for every $\mathbf{z}$:

$$\mathbb{E}[\|g(x, U_{\mathbf{z}}) - \nabla F_{\mathbf{z}}(x)\|^2] \leq 2\delta(M^2\|x\|^2 + B^2)\,.$$

Lyapunov function for underdamped dynamics

(v) The law μ_0 of the initial state (X_0, V_0) of SGHMC satisfies:

$$\int_{\mathbb{R}^{2d}} e^{\alpha \mathcal{V}(x,v)} \mu_0(dx, dv) < \infty \,,$$

where $\mathcal{V}$ is a Lyapunov function:

$$\mathcal{V}(x,v) := \beta F_{\mathbf{z}}(x) + \frac{\beta}{4}\gamma^2 (\|x + \gamma^{-1}v\|^2 + \|\gamma^{-1}v\|^2 - \lambda \|x\|^2) \,, \tag{11}$$

and α is a positive explicit constant and λ is a positive constant less than $\min(1/4, m/(M + \gamma^2/2))$.

- The Lyapunov function $\mathcal{V}$ is used in Eberle et al. (2019) to study the rate of convergence to equilibrium for underdamped Langevin diffusion.

Main Result

Theorem (Gao, Gürbüzbalaban and Zhu (2021+))

Consider the SGHMC2 iterates $(\hat{X}_k, \hat{V}_k)$. If Assumptions (i)-(v) are satisfied, then for $\beta, \varepsilon > 0$, we have

$$\left|\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X))\right| \leq \mathcal{J}_0(\mathbf{z},\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon),$$

provided that

$$\eta \leq \min\left\{\frac{\varepsilon^2}{\log(1/\varepsilon)}, \quad Constant(d,\beta)\right\}, \tag{12}$$

and

$$k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \geq e. \tag{13}$$

Formulas and interpretations of the upper bounds

- The parameter μ_* governs the speed of convergence to the equilibrium of the continuous-time underdamped Langevin diffusion (Eberle et al. (2019)).
- $\mathcal{J}_0(\mathbf{z}, \varepsilon)$ quantifies the dependency on the initialization μ_0 and the dataset $\mathbf{z}$.

$$\mathcal{J}_0(\mathbf{z}, \varepsilon) := Const \cdot \sqrt{\mathcal{H}_\rho(\mu_0, \pi_{\mathbf{z}})} \cdot \varepsilon \leq \overline{\mathcal{J}}_0(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{d+\beta}{\mu_* \beta^{3/4}}\varepsilon\right),$$

- $\hat{\mathcal{J}}_1(\varepsilon)$ is controlled by the discretization error and the amount of noise parameter δ in the gradients.

$$\hat{\mathcal{J}}_1(\varepsilon) = \tilde{\mathcal{O}}\left(\frac{(d+\beta)^{3/2}}{\beta\sqrt{\mu_*}}\left(\sqrt{\log(\varepsilon^{-1})}\delta^{1/4} + \varepsilon\right)\sqrt{\log(\log(\varepsilon^{-1})/\mu_*)}\right).$$

Performance Bound for Empirical Risk Minimization

- Note that the expected excess empirical risk can be decomposed:

$$\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) = \mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X)) + \mathbb{E}_{(X,V)\sim\pi_{\mathbf{z}}}(F_{\mathbf{z}}(X)) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x)$$

- For finite β, one can derive (Raginsky et al. (2017))

$$\int_{\mathbb{R}^{2d}} F_{\mathbf{z}}(x)\pi_{\mathbf{z}}(dx,dv) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \leq \mathcal{J}_2 := \frac{d}{2\beta}\log\left(\frac{eM(\frac{b\beta}{d}+1)}{m}\right).$$

  - x-marginal of $\pi_{\mathbf{z}}(dx,dv)$ is the same as the stationary distribution of the overdamped Langevin SDE.

Performance Bound for Empirical Risk Minimization

Corollary (Gao, Gürbüzbalaban and Zhu (2021+))

Under the setting of Theorem 1, the empirical risk minimization problem admits the performance bounds:

$$\mathbb{E}F_{\mathbf{z}}(\hat{X}_k) - \min_{x\in\mathbb{R}^d} F_{\mathbf{z}}(x) \leq \mathcal{J}_0(\varepsilon, \mathbf{z}) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2\,, \tag{14}$$

provided that

$$\eta \leq \min\left\{\frac{\varepsilon^2}{\log(1/\varepsilon)}, \quad Constant(d, \beta)\right\},$$

and $k\eta = \frac{1}{\mu_*}\log\left(\frac{1}{\varepsilon}\right) \geq e$.

Performance bound for Population Risk Minimization

Corollary (Gao, Gürbüzbalaban and Zhu (2021+))

Under the setting of Theorem 1, the expected excess risk of $\hat{X}_k$ is bounded by

$$\mathbb{E}F(\hat{X}_k) - F^* \leq \overline{\mathcal{J}}_0(\varepsilon) + \hat{\mathcal{J}}_1(\varepsilon) + \mathcal{J}_2 + \mathcal{J}_3(n)\,,$$

with

$$\mathcal{J}_3(n) := \frac{4\beta c_{LS}}{n}\left(\frac{M^2}{m}(b + d/\beta) + B^2\right), \tag{15}$$

where c_{LS} is a constant that can be upper bounded.

- $\mathcal{J}_3(n)$ controls the difference between the finite sample size problem (3) and the original problem (1).

Performance comparison with respect to SGLD algorithm

- For the expected empirical risk $\tilde{\mathcal{O}}(\hat{\varepsilon})$, we have

$$K_{SGHMC2} = \tilde{\Omega}\left(\frac{d}{\mu_*^3\hat{\varepsilon}^3}\right), \qquad \hat{K}_{SGHMC2} = \tilde{\Omega}\left(\frac{d^3}{\mu_*^5\hat{\varepsilon}^9}\right),$$

$$K_{SGLD} = \tilde{\Omega}\left(\frac{d^{14}}{\lambda_*^5\hat{\varepsilon}^{18}}\right), \qquad \hat{K}_{SGLD} = \tilde{\Omega}\left(\frac{d^{26}}{\lambda_*^9\hat{\varepsilon}^{34}}\right),$$

- K denotes the number of iterates and $\hat{K}$ denotes the stochastic gradient computations, defined as $\hat{K} = K\delta^{-1}$, since δ^{-1} can be interpreted as mini-batch size.
- λ_* is the uniform spectral gap for the continuous-time overdamped Langevin diffusion (Raginsky et al. (2017)).

Examples of non-convex functions

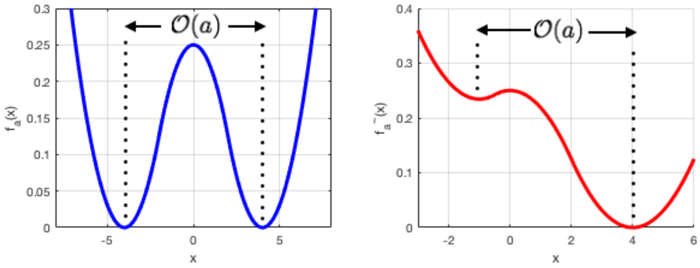

Figure: The illustration of the functions $f_a(x)$ (left) and $\tilde{f}_a(x)$ (right) for $a = 4$.

Comparison of λ_* and μ_*

Proposition (Gao, Gürbüzbalaban and Zhu (2021+))

Under certain conditions,

$$\lambda_* = \tilde{\mathcal{O}}(a^{-2}), \qquad \mu_* = \Theta(a^{-1}).$$

- Parameters λ_* and μ_* govern the convergence rate to the equilibrium of the overdamped and underdamped Langevin SDE; $\frac{1}{\lambda_*}$ and $\frac{1}{\mu_*}$ can be both exponentially large in dimension and β.
- Since under certain conditions $\frac{1}{\mu_*} = \mathcal{O}\left(\sqrt{\frac{1}{\lambda_*}}\right)$, if the other parameters (β, d, δ) are fixed, since under many examples, then SGHMC can lead to an improvement upon the SGLD performance.

Conclusion

- SGHMC is a momentum-based popular variant of stochastic gradient where a controlled amount of Gaussian noise is added to the gradient estimates for optimizing a non-convex function.
- We obtained first-time finite-time guarantees for the convergence of SGHMC to the ε-global minimizers under some regularity assumption on the non-convex objective f.
- We also show that on a class of non-convex problems, SGHMC can be faster than overdamped Langevin MCMC approaches such as SGLD.
- Our results show that momentum-based acceleration is possible on a class of non-convex problems under some conditions.

Further Related Works

- Breaking reversibility accelerates non-convex optimization for Langevin algorithms [6].
- Heavy-tailed Langevin dynamics with α-stable Lévy noise [7].
- Decentralized SGLD and SGHMC [8].

---

[6] Gao, X., Gürbüzbalaban, M. and L. Zhu (2020). Breaking reversibility accelerates Langevin dynamics for global non-convex optimization. *Advances in Neural Information Processing Systems* **33** (NeurIPS 2020).

[7] Şimşekli, U., Zhu, L., Teh, Y. and M. Gürbüzbalaban (2020). Fractional underdamped Langevin dynamics: Retargeting SGD with momentum under heavy-tailed gradient noise. *International Conference on Machine Learning*.

[8] Gürbüzbalaban, M., Gao, X., Hu, Y. and L. Zhu (2021). Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*. **22**, 1-69.

Thank you

Thank you! Questions?