Stochastic Approximate Gradient Descent via the Underdamped Langevin Algorithm

Carnegie Mellon University

Yixuan Qiu
Carnegie Mellon University

Xiao Wang
Purdue University

PURDUE UNIVERSITY

Introduction

- The stochastic gradient descent method (SGD) is one of the most popular and widely-used optimization techniques in large-scale machine learning problems.
- For a broad range of problems, the objective function is an expectation, $F(\theta) = \mathbb{E}[f(\theta;\xi)]$.
- We consider the case that ξ follows a complicated continuous distribution $\pi(\xi)$, so that an unbiased stochastic gradient cannot be trivially obtained.
- We propose the stochastic *approximate* gradient descent (SAGD) as an extension to SGD for such cases, based on the underdamped Langevin sampling algorithm.
- Theoretical and empirical studies demonstrate the validity and usefulness of SAGD.

The Underdamped Langevin Algorithm

- We use a stochastic gradient $\tilde{g}(\theta) = K^{-1}\sum_{k=0}^{K-1}\nabla f(\theta;\xi_k)$ to approximate the true gradient $g(\theta) = \mathbb{E}[\nabla f(\theta;\xi)]$, where $\{\xi_k\}$ is generated by the underdamped Langevin algorithm.
- To sample from a multi-dimensional distribution with density $\pi(\xi) \propto \exp\{-V(\xi)\}$:

$$\xi_{k+1} = \xi_k + \delta\rho_k,$$
$$\rho_{k+1} = (1-\gamma\delta)\rho_k - \delta\cdot\nabla V(\xi_k) + \sqrt{2\gamma\delta}\eta_k, \quad k \geq 0,$$

  where δ is the step size, $\{\eta_k\} \overset{iid}{\sim} N(0, I_r)$, and η_k is independent of $\{(\xi_k, \rho_k)\}_{i=0}^{k-1}$.
- γ, ξ_0, and ρ_0 are arbitrary constants.

Theorem

Under regularity conditions, there exist constants $C_1 > 0$ and $C_2 > 0$ such that for any $\gamma > 0$ and $K > 0$, we have

$$|\mathbb{E}(\tilde{g}) - g| \leq C_1\left(\frac{1}{K\delta} + \delta\right), \qquad \mathbb{E}\left[(\tilde{g} - g)^2\right] \leq C_2\left(\frac{1}{K\delta} + \delta^2\right).$$

Stochastic Approximate Gradient Descent

- Outline of the SAGD algorithm.

SAGD for minimizing $F(\theta) = \mathbb{E}[f(\theta;\xi)]$

1. **For** $t = 0, 1, \ldots, T-1$ **Do**
2. $\quad \xi_{t,0} \leftarrow \xi_0$, $\rho_{t,0} \leftarrow \rho_0$
3. $\quad$ **For** $k = 1, \ldots, K_t - 1$ **Do**
4. $\quad\quad \xi_{t,k+1} \leftarrow \xi_{t,k} + \delta_t\rho_{t,k}$
5. $\quad\quad \rho_{t,k+1} \leftarrow (1-\gamma\delta_t)\rho_{t,k} - \delta_t\cdot\nabla V(\xi_{t,k}) + \sqrt{2\gamma\delta_t}\eta_{t,k}$, where $\eta_{t,k} \sim N(0, I_r)$
6. $\quad$ **End For**
7. $\quad \tilde{g}_t(\theta) \leftarrow K_t^{-1}\sum_{k=0}^{K_t-1}\nabla f(\theta;\xi_{t,k})$
8. $\quad \theta_{t+1} \leftarrow \mathcal{P}_\Theta\left(\theta_t - \alpha_t\cdot\tilde{g}_t(\theta_t)\right)$
9. **End For**

- Convergence guarantee for convex objective functions.

Theorem

Suppose that $F(\theta)$ is convex and L-Lipschitz continuous in $\theta \in \Theta$, and Θ is a closed convex set with diameter $D < \infty$. Choose $\delta_t = C_1/\sqrt{t}$, $K_t = C_2 t$, and $\alpha_t = \alpha_0/\sqrt{t}$, where $C_1, C_2, \alpha_0 > 0$ are constants. Then under regularity conditions, we have

$$\mathbb{E}[F(\hat{\theta})] - F^* \leq \mathcal{O}(1/\sqrt{T}).$$

- Convergence guarantee for non-convex objective functions.

Theorem

Suppose that $g(\theta)$ is G-Lipschitz continuous in θ, and let $\delta_t = C_1 t^{-c}$, $K_t = C_2 t^{2c}$, and $\alpha_t = \alpha_0/t$ for some constants $0 < \alpha_0 < 1/(2G)$ and $C_1, C_2, c > 0$. Then under regularity conditions, we have

$$\liminf_{t\to\infty}\ \mathbb{E}[\|g(\theta_t)\|^2] = 0.$$

Application I - Automated EM Algorithm

- Latent variables $Z_1, \ldots, Z_n \overset{iid}{\sim} N(0,1)$ and data $X_i|\{Z_1 = z_1, \ldots, Z_n = z_n\} \sim \mathsf{Gamma}(10\cdot\sigma(a+bz_i)), \sigma(x) = 1/(1+\exp(-x))$. True parameters $\theta = (a,b) = (2, 0.5)$.
- Use EM algorithm to estimate θ: $\theta_{k+1} = \arg\max_\theta Q(\theta;\theta_k) = \mathbb{E}_{Z|X=x,\theta_k}[L(\theta;x,Z)]$.
- In each M-step, we run SAGD and exact gradient descent for $T = 100$ iterations.

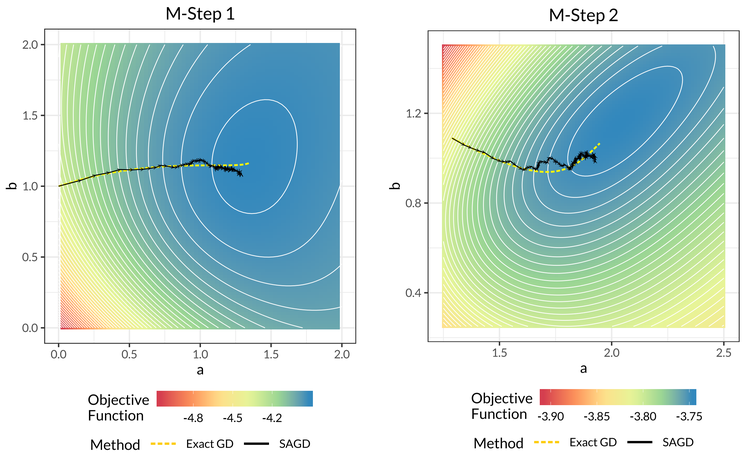

Figure 1: Comparison of exact GD and SAGD for two M-steps. The path of (a, b) values are overlaid on the contour plot of the $Q(\theta;\theta_k)$ function.

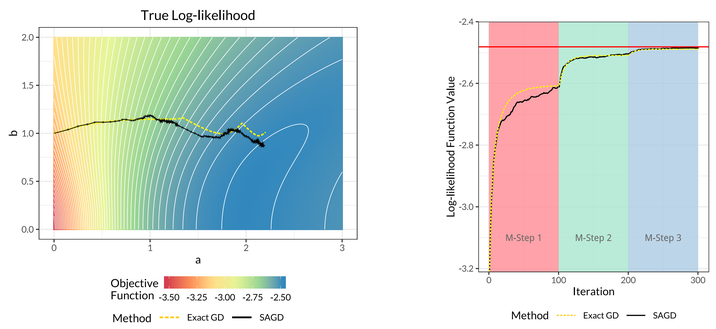

Figure 2: Left: The paths of (a, b) on the surface of the true log-likelihood function. Right: The log-likelihood function value versus the number of gradient updates.

Application II - Bias correction for VAE

- We show how SAGD can be used to correct the bias of VAE for distribution matching.
- First, generate a latent random vector $U = (U_1, \ldots, U_{10})^{\mathrm{T}}$ with $P(U_1 \leq u_1, \ldots, U_{10} \leq u_{10}) = C(F_1(u_1), \ldots, F_{10}(u_{10}))$, where $C(v_1, \ldots, v_{10}) = \varphi^{-1}(\varphi(v_1) + \cdots + \varphi(v_{10}))$, $\varphi(t) = t^{-2} - 1$, and F_i is the distribution function of $\mathsf{Gamma}(1 + (i-1)/9)$.
- Given the latent variable U, the data point X is simulated as $X|\{U = u\} \sim N(u, I_{10})$.
- Create a simulated dat set with sample size $n = 10000$.
- Consider a VAE model by assuming $U = h(Z)$, where $Z \sim N(0, I_{10})$ and $h : \mathbb{R}^{10} \mapsto \mathbb{R}^{10}$ is a deep neural network (DNN) transformation.
- First train VAE on the data set, and then use SAGD to refine the parameters of h.
- Compare the estimated distribution of U by VAE and SAGD.

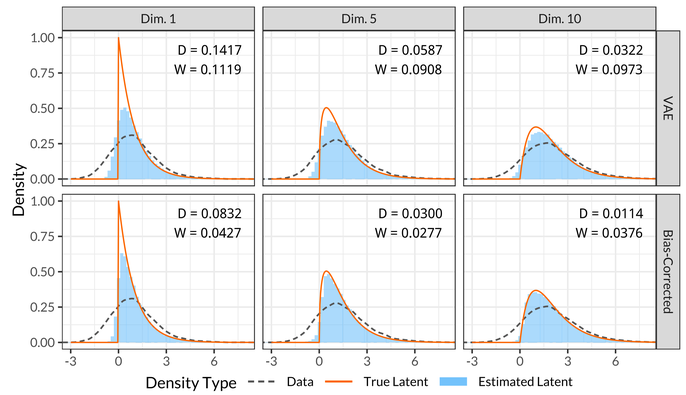

Figure 3: A demonstration of the marginal data distribution, true marginal latent density, and estimated marginal latent distributions. D and W stand for the K–S distance and 1-Wasserstein distance, respectively.

Conclusion

- SAGD is an extension to SGD for complicated statistical and machine learning problems.
- We prove the convergence of SAGD for both convex and non-convex objective functions.
- SAGD can be applied to important statistical and machine learning problems such as the EM algorithm and VAE.

Selected References

- Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 22(3):400–407
- Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Review*, 60(2), 223-311.
- Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th ICML*, 681–688.
- Kingma, D. P. and Welling, M. (2014). Stochastic gradient VB and the variational auto-encoder. In *Proceedings of the 2nd ICLR*.

All correspondence goes to Yixuan Qiu <yixuanq@andrew.cmu.edu>