A Latent Variational Framework for Stochastic Optimization

Philippe Casgrain [1,2]

[1]University of Toronto [2]Citadel LLC.

Abstract

This paper provides a unifying theoretical framework for stochastic optimization algorithms by means of a latent stochastic variational problem. Using techniques from stochastic control, the solution to the variational problem is shown to be equivalent to that of a Forward Backward Stochastic Differential Equation (FBSDE). By solving these equations, we recover a variety of existing adaptive stochastic gradient descent methods. This framework establishes a direct connection between stochastic optimization algorithms and a secondary latent inference problem on gradients, where a prior measure on noisy gradient observations determine the resulting algorithm.

Introduction

- Stochastic optimization algorithms are important tools in machine learning, particularly in the optimization problems that arise from deep learning.
- Stochastic optimization algorithms overcome the computational hurdles of large scale optimization problems **by replacing the exact computation of the gradients with more easily computable statistical samples**.
- There exists a wide variety of stochastic optimization algorithms (see e.g. [1, 2, 3, 5, 6, 8, 9, 12]), **yet no single theoretical interpretation** with which they can all be understood and compared.
- This paper **proposes a theoretical model for stochastic optimization** using a continuous-time SDE-based interpretation of these algorithms, similar to those used in [4, 7, 10, 11].
- The approach taken here is in the same spirit as that found in [10] for deterministic optimization.
- We construct a variational problem to model the task of **selecting optimization algorithms which maximize their average performance** over a set of optimization tasks. **Stochastic optimization algorithms** naturally emerge as solutions to the variational problem.

Stochastic Optimization

- The objective is to construct an algorithm $\{x_{t_k}\}$ which can **minimize a risk function** $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}$ of the form
$$f(x) = \frac{1}{|\mathfrak{N}|} \sum_{z \in \mathfrak{N}} \ell(x; z) \;, \tag{1}$$
where $\ell : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, and $\mathfrak{N} := \{z_i \in \mathcal{Z} \;,\; i = 1, \dots, N\}$ is a set of training points.
- We typically assume that N and d are large, so that **computing $\boldsymbol{\nabla f}$ is a computationally expensive operation**.
- At iteration k, rather than computing $\nabla f(x_{t_k})$, we collect **computationally cheap** noisy gradient samples,
$$g_{t_k} = \frac{1}{|\mathfrak{N}^m_{t_k}|} \sum_{z \in \mathfrak{N}^m_{t_k}} \nabla_x \ell(x_{t_k}; z) \;, \tag{2}$$
where for each t, $\mathfrak{N}^m_t \subseteq \mathfrak{N}$ is an independent sample of size $m \le N$ from the set of training points.
- Stochastic optimization algorithms then use the noisy gradient samples to minimize f.
- At each iteration K, stochastic optimization algorithms are restricted so that they may only use the collection $\{g_{t_k}\}_{k=1}^K$ of past noisy gradients to compute the next step x_{t_K}.
- The key property is that these algorithms can approximate $\boldsymbol{x^* = \arg\min_x f(x)}$, without having to directly observe its gradients.

Continuous-Time Optimization

We approximate the above with a continuous-time model.

- Given a randomly generated collection of optimization problems, and a fixed run-time $T > 0$, we wish to **determine which algorithms $\boldsymbol{X}$ achieve optimal performance on average**.
- To model the collection of optimization problems, we define a random objective function $\boldsymbol{f(x)}$, satisfying $f : \mathbb{R}^d \to \mathbb{R}$ and $f \in C^2(\mathbb{R}^d)$ almost surely.
- Each draw from this random variable generates a new optimization problem: $\boldsymbol{f(x) \to \min}$.
- We define an algorithm $X = (X_t)_{t \ge 0}$ as a differentiable, continuous-time process satisfying $X_t \in \mathbb{R}^d$ and $\frac{dX}{dt} = \dot{X}_t$. We can interpret this model as an algorithm taking steps
$$X_{t+\epsilon} \approx X_t + \epsilon \, \dot{X}_t \;,$$
over the short time intervals $[t, t + \epsilon]$.
- As we optimize, we collect observations from a noisy gradient process $\boldsymbol{(g_t)_{t \ge 0}}$. This process can be seen as a continuous-time interpretation of equation (2).
- To preserve the stochastic algorithm's information restriction, we only consider algorithms $\boldsymbol{X}$, which are adapted to the filtration $\boldsymbol{\mathcal{F}_t := \sigma\left((g_u)_{0 \le u \le t}\right)}$ generated by the noisy gradients.
- This condition restricts $\boldsymbol{X}$ so that it may only use information from $\boldsymbol{g}$, and cannot directly observe $\boldsymbol{f}$ or $\boldsymbol{\nabla f}$.

The Variational Problem

- We define an objective functional $\mathcal{J}$ as
$$\mathcal{J}(X) = \mathbb{E}\left[\boldsymbol{f(X_T)} + e^{-\delta_T} \int_0^T \boldsymbol{\mathcal{L}(t, X_t, \dot{X}_t)\, dt} \right] \;, \tag{3}$$
$$\mathcal{L}(t, X, \nu) = e^{\gamma_t}\left(e^{\alpha_t} D_h\left(X + e^{-\alpha_t}\nu, X\right) - e^{\beta_t} f(X)\right) \;,$$
where $\alpha, \beta, \delta : \mathbb{R}^+ \to \mathbb{R}$ are continuously differentiable functions and where D_h is the Bregman divergence, $D_h(X, Y) = h(X) - h(Y) - \langle \nabla h(Y), Y - X \rangle$ for $h \in C^2(\mathbb{R}^d)$ and strictly convex.
- We interpret the objective (3) to represent the sum of
  1. The algorithm's average performance after a fixed run-time $\boldsymbol{T}$, $\mathbb{E}[\boldsymbol{f(X_T)}]$.
  2. A regularization term, which penalizes the total pathwise 'energy' spent by the algorithm to reach $\boldsymbol{X_T}$.
- We seek an algorithm $\boldsymbol{X^\star}$ such that $\boldsymbol{X^\star = \arg\min_{X \in \mathcal{A}} \mathcal{J}(X)}$, where $\mathcal{A}$ is the collection of $\mathcal{F}_t$-adapted processes.
- **This is a latent control problem**, since X cannot directly observe the loss function, f.

Main Results

Applying techniques from the calculus of variations, **we arrive at optimality conditions for the variational problem**, as well as **rates of convergence for the optimal algorithm**.

Theorem 1. (Solution to the Variational Problem) An algorithm $\boldsymbol{X}$ is a critical point of $\mathcal{J}$ if and only if the FBSDE
$$\begin{aligned} d\left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_t &= \mathbb{E}\left[\left(\frac{\partial \mathcal{L}}{\partial X}\right)_t \middle| \mathcal{F}_t\right] dt + d\mathcal{M}_t \quad \forall t < T \,, \\ \left(\frac{\partial \mathcal{L}}{\partial \nu}\right)_T &= -e^{\delta_T}\, \mathbb{E}\left[\nabla f(X_T) \middle| \mathcal{F}_T\right] \end{aligned} \tag{4}$$
holds, where $\mathcal{M} = (\mathcal{M}_t)_{0 \le t \le T}$ is an $\mathcal{F}_t$-adapted martingale.

Theorem 2. (Rate of Convergence) Assume that f is almost surely convex, $\dot{\gamma}_t = e^{\alpha_t}$ and $\dot{\beta}_t \le e^{\alpha_t}$. Moreover, assume that h is L-Lipschitz smooth and μ-strongly-convex. Define $x^\star$ to be a global minimum of f. If $x^\star$ exists almost surely, the optimizer defined by FBSDE (4) satisfies
$$\mathbb{E}\left[f(X_t) - f(x^\star)\right] = O\left(e^{-\beta_t} \max\left\{1 \,, e^{-2\gamma_t}\, \mathbb{E}\left[\, [\mathcal{M}]_t \,\right]\right\}\right) \;, \tag{5}$$
where $[\mathcal{M}]_t$ is the quadratic variation of the process $\mathcal{M}_t$.

Connection to Discrete Algorithms

Using the optimality equation (4), **we can recover a variety of well-known optimization algorithms** by specifying various models for loss functions and gradients. The steps we take are as follows:

1. **Specify a model** for the gradients of the loss function, $\boldsymbol{(\nabla f(X_t))_{t \ge 0}}$, and a model for the noisy observations of these gradients, $\boldsymbol{(g_t)_{t \ge 0}}$.
2. **Solve the optimality equation (4)** directly, or approximate the solution using a singular perturbation technique.
3. **Discretize the continuous solution** over the finite mesh
$$\mathcal{T} = \{t_0 = 0 \,,\; t_{k+1} = t_k + e^{-\alpha_{t_k}} : k \in \mathbb{N}\}$$
to obtain a discrete optimization algorithm.

Stochastic Mirror Descent & Stochastic Gradient Descent

- The model:
  - Assume that gradients evolve as $\boldsymbol{\nabla f(X_t) = \sigma W^f_t}$ and that noisy gradients are sampled according to $\boldsymbol{g_t = \sigma(W^f_t + \rho\, W^e_t)}$.
  - We assume that $\sigma, \rho > 0$ and $(W^e_t, W^f_t)_{t \ge 0}$ are independent Brownian motions of size d.
- Solving and discretizing equation (4) gives the update rule
$$X_{t_{k+1}} = \nabla h^*\left(\nabla h(X_{t_k}) - \tilde{\Phi}_{t_k}\, g_{t_k}\right) \;,$$
where $\tilde{\Phi}_t$ is a time-dependent learning rate.
- This algorithm **corresponds exactly to stochastic mirror descent**, where the special case of $h(x) = \frac{1}{2}\|x\|^2$ recovers stochastic gradient descent.
- We can interpret result as showing that gradient descent implicitly assumes **that true gradients and the noise in stochastic gradients are martingales**. SGD/mirror descent are optimal when gradient evolution is structureless and gradients are sampled with IID noise.

Kalman Gradient Descent & Stochastic Momentum Descent

- The model:
  - We assume that gradients take the form $\boldsymbol{\nabla f(X_t) = b^\intercal y_t}$, where $y_t \in \mathbb{R}^k$ evolves according to the dynamics $\boldsymbol{dy_t = -Ay_t\, dt + BdW_t}$.
  - Noisy gradients are observed according to $\boldsymbol{dg_t = \nabla f(X_t)\, dt + \sigma dB_t}$.
  - $b \in \mathbb{R}^{k \times d}$, $A, B, \sigma \in \mathbb{R}^{k \times k}$ are nonnegative-definite, $(W_t, B_t)_{t \ge 0}$ are indep. Brownian Motions of size k and d.
- This model generates the update rule
$$X_{t_{k+1}} = \nabla h^*\left(\nabla h(X_{t_k}) - b^\intercal \tilde{\Phi}_{t_k} \hat{y}_{t_k}\right) \;,$$
where $\boldsymbol{\hat{y}_t}$ **is the Kalman filter of $\boldsymbol{y_t}$**, which satisfies $b^\intercal \hat{y}_t = \mathbb{E}\left[\nabla f(X_t) \mid \{g_{t_{k'}}\}_{k' \le k}\right]$, and $\tilde{\Phi}_t \in \mathbb{R}^{d \times k}$ is a deterministic function of time.
- When $h(x) = \frac{1}{2}\|x\|^2$, this corresponds exactly to Kalman Gradient Descent of [9].
- Letting $k \to \infty$ and $h(x) = \frac{1}{2}\|x\|^2$, we find that the asymptotic update rule takes the form $X_{t_{k+1}} = \Psi^{(0)}_{t_k} X_{t_k} + \Psi^{(1)}_{t_k} g_{t_k}$, where $\Psi^{(0)}_{t_k}, \Psi^{(1)}_{t_k}$ are time-dependent matrices, **which corresponds to stochastic momentum descent** with time-varying learning and decay rates.
- This demonstrates that Kalman Gradient Descent and Stochastic Momentum Descent are in fact related algorithms, and that they are optimal when **gradients are expected to decay exponentially in time and stochastic gradient noise is IID**.

Conclusion

- We constructed a model which **captures the latent elements of stochastic optimization** within a variational problem.
- We identified the optimal solution to this problem in the form of an FBSDE.
- Using this model, we **identified the circumstances in which various stochastic optimization algorithms are optimal**.

References

[1] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[2] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] W. Krichene, A. Bayen, and P. L. Bartlett. Accelerated mirror descent in continuous and discrete time. In *Advances in neural information processing systems*, pages 2845–2853, 2015.

[5] A. S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[6] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/k^ 2). In *Doklady AN USSR*, volume 269, pages 543–547, 1983.

[7] M. Raginsky and J. Bouvrie. Continuous-time stochastic mirror descent on a network: Variance reduction, consensus, convergence. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 6793–6800. IEEE, 2012.

[8] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[9] J. Vuckovic. Kalman gradient descent: Adaptive variance reduction in stochastic optimization. *arXiv preprint arXiv:1810.12273*, 2018.

[10] A. Wibisono, A. C. Wilson, and M. I. Jordan. A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358, 2016.

[11] P. Xu, T. Wang, and Q. Gu. Continuous and discrete-time accelerated stochastic mirror descent for strongly convex functions. In *International Conference on Machine Learning*, pages 5488–5497, 2018.

[12] M. D. Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.