Diffusion Approximations and Applications in Nonconvex Optimization

Binan Gu

Department of Mathematical Sciences, New Jersey Institute of Technology

New Jersey Institute of Technology
Spring 2021 Machine Learning Talk VIII

Main scientists

Monroe D. Donsker (Mathematician)

- Courant Institute

Main scientists

Monroe D. Donsker (Mathematician)

- Courant Institute
- Large Deviation Theory with S.R.S. Varadhan

Main scientists

Monroe D. Donsker (Mathematician)

- Courant Institute
- Large Deviation Theory with S.R.S. Varadhan
- For today's work: Donsker's theorem, i.e. Functional Central Limit Theorems.

Peter W. Glynn (Applied Mathematician)

Main scientists

Monroe D. Donsker (Mathematician)

- Courant Institute
- Large Deviation Theory with S.R.S. Varadhan
- For today's work: Donsker's theorem, i.e. Functional Central Limit Theorems.

Peter W. Glynn (Applied Mathematician)

- Professor at Stanford, Operations Research

Main scientists

Monroe D. Donsker (Mathematician)

- Courant Institute
- Large Deviation Theory with S.R.S. Varadhan
- For today's work: Donsker's theorem, i.e. Functional Central Limit Theorems.

Peter W. Glynn (Applied Mathematician)

- Professor at Stanford, Operations Research
- Former chair of Department of Management Science and Engineering at Stanford.
- For today's work: Diffusion approximations for complicated stochastic processes.

Motivation

Nonconvex Optimization

We aim to solve the following nonconvex optimization problem:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x}; \xi)\right]$$

where ξ is sampled from some distribution $\mathcal{D}$.

Motivation

Nonconvex Optimization

We aim to solve the following nonconvex optimization problem:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x}; \xi) \right]$$

where ξ is sampled from some distribution $\mathcal{D}$.

Stochastic Gradient Descent and its Variants

We evaluate the noisy gradient and update by

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k \right)$$

where η is step-size/learning rate, and noise ξ_k independent of the sigma algebra generated up to $\mathbf{x}^{(k-1)}$.

Motivation

Nonconvex Optimization

We aim to solve the following nonconvex optimization problem:

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x}; \xi)\right]$$

where ξ is sampled from some distribution $\mathcal{D}$.

Stochastic Gradient Descent and its Variants

We evaluate the noisy gradient and update by

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right)$$

where η is step-size/learning rate, and noise ξ_k independent of the sigma algebra generated up to $\mathbf{x}^{(k-1)}$.

Notice that $\mathbf{x}^{(k)}$ forms a discrete-time Markov process with laws determined by ξ (and hence time-homogeneous).

Diffusion Approximations

Rescaled Random Walk

Consider $X_1, X_2, \ldots$ i.i.d. centered random variables with unit variance. Write $S_n = \sum_{i=1}^{n} X_i$.

Diffusion Approximations

Rescaled Random Walk

Consider $X_1, X_2, \ldots$ i.i.d. centered random variables with unit variance. Write $S_n = \sum_{i=1}^{n} X_i$. Then

$$W^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow{d} W(t) \sim \mathcal{N}(0, t).$$

Diffusion Approximations

Rescaled Random Walk

Consider $X_1, X_2, \ldots$ i.i.d. centered random variables with unit variance. Write $S_n = \sum_{i=1}^{n} X_i$. Then

$$W^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow{d} W(t) \sim \mathcal{N}(0, t).$$

Note that $W^{(n)}(1) \xrightarrow{d} \mathcal{N}(0, 1)$ by classical CLT.

Diffusion Approximations

Rescaled Random Walk

Consider $X_1, X_2, \ldots$ i.i.d. centered random variables with unit variance. Write $S_n = \sum_{i=1}^{n} X_i$. Then

$$W^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow{d} W(t) \sim \mathcal{N}(0, t).$$

Note that $W^{(n)}(1) \xrightarrow{d} \mathcal{N}(0,1)$ by classical CLT.

Donsker's Theorem, Functional CLT

Consider $X_1, X_2, \ldots$ i.i.d with distribution F. Define the empirical distribution function $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i < x}$.

Diffusion Approximations

Rescaled Random Walk

Consider $X_1, X_2, \ldots$ i.i.d. centered random variables with unit variance. Write $S_n = \sum_{i=1}^n X_i$. Then

$$W^{(n)}(t) := \frac{S_{\lfloor nt \rfloor}}{\sqrt{n}} \xrightarrow{d} W(t) \sim \mathcal{N}(0, t).$$

Note that $W^{(n)}(1) \xrightarrow{d} \mathcal{N}(0, 1)$ by classical CLT.

Donsker's Theorem, Functional CLT

Consider $X_1, X_2, \ldots$ i.i.d with distribution F. Define the empirical distribution function $F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{X_i < x}$. Then

$$G_n(x) = \sqrt{n}\left(F_n(x) - F(x)\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_t)$$

where the covariance Σ_t

$$\mathrm{Cov}\left(G(s), G(t)\right) = \min\left\{F(s), F(t)\right\} - F(s) F(t).$$

Take-aways from FCLT

- Discrete processes, properly scaled, can be approximated by Brownian motions weakly.

Take-aways from FCLT

- Discrete processes, properly scaled, can be approximated by Brownian motions weakly. For example, the rescaled random walk (mean μ and covariance Σ in general) satisfies
$$S_{\lfloor nt \rfloor} \stackrel{d}{\approx} \mu nt + \sqrt{n\Sigma} B(t)$$
- Continuous mapping principle applies (e.g. $h(X_t)$ where h is continuous).

Take-aways from FCLT

- Discrete processes, properly scaled, can be approximated by Brownian motions weakly. For example, the rescaled random walk (mean μ and covariance Σ in general) satisfies
$$S_{\lfloor nt \rfloor} \overset{d}{\approx} \mu nt + \sqrt{n\Sigma} B(t)$$
- Continuous mapping principle applies (e.g. $h(X_t)$ where h is continuous).
- Identify scaled means and covariances to derive an SDE for the approximant.

Pros and Cons

Why approximations?

- Diffusion is easier to study than discrete processes.

Pros and Cons

Why approximations?

- Diffusion is easier to study than discrete processes.
- One can prove concentration results to learn about convergence rates, e.g. Komlós–Major–Tusnády approximation [6] that improves Donsker's theorem.

Pros and Cons

Why approximations?

- Diffusion is easier to study than discrete processes.
- One can prove concentration results to learn about convergence rates, e.g. Komlós–Major–Tusnády approximation [6] that improves Donsker's theorem.
- For variants of SGD, only the functional form changes. The approximation approach is general. (see Momentum SGD as another example [5]).

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right) + \mu\left(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\right)$$

(This gives rise to an Ornstein-Uhlenbeck process which has a closed form analytical solution.)

Pros and Cons

Why approximations?

- Diffusion is easier to study than discrete processes.
- One can prove concentration results to learn about convergence rates, e.g. Komlós–Major–Tusnády approximation [6] that improves Donsker's theorem.
- For variants of SGD, only the functional form changes. The approximation approach is general. (see Momentum SGD as another example [5]).

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right) + \mu\left(\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\right)$$

(This gives rise to an Ornstein-Uhlenbeck process which has a closed form analytical solution.)

Questions to address

Convergence, rate of convergence, accuracy (in what sense?)

Diffusion Approximations of SGD

$$\min_{\mathbf{x}} F(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x};\xi)\right]$$

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)};\xi_k\right)$$

[3] showed that the discrete Markov process $\mathbf{x}^{(k)}$ can be approximated (in the sense of weak accuracy) by the scaled solution $\boldsymbol{X}(k\eta)$ to the SDE (for finite time $[0, T]$)

$$d\boldsymbol{X}(s) = b\left(\boldsymbol{X}(s)\right) ds + \sqrt{\eta}\mathbf{S}\left(\boldsymbol{X}(s)\right) d\boldsymbol{B}(s), \quad \boldsymbol{X}(0) = \mathbf{x}^{(0)}$$

$$b(\mathbf{x}) = -\nabla F(\mathbf{x}) - \frac{1}{4}\eta\nabla\left|\nabla F(\mathbf{x})\right|^2$$

$$\mathbf{S}(\mathbf{x}) = \sqrt{\mathrm{var}\left(\nabla f(\mathbf{x};\xi)\right)}$$

How to get this SDE?

Basic technique [1, 2] to turn the stochastic algorithm around some local optimum $\mathbf{x}^*$ into some SDE

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right)$$

1. Work with error terms, i.e. normalize the variable by

$$\mathbf{u}_\eta^{(k)} = \frac{\mathbf{x}_\eta^{(k)} - \mathbf{x}^*}{\sqrt{\eta}}, \textit{ as } \mathrm{Var}\left(\mathbf{x}_{\lfloor \frac{1}{\eta} \rfloor} - \mathbf{x}^*\right) = O(\eta).$$

How to get this SDE?

Basic technique [1, 2] to turn the stochastic algorithm around some local optimum $\mathbf{x}^*$ into some SDE

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right)$$

1. Work with error terms, i.e. normalize the variable by

$$\mathbf{u}_\eta^{(k)} = \frac{\mathbf{x}_\eta^{(k)} - \mathbf{x}^*}{\sqrt{\eta}}, \textit{ as } \mathrm{Var}\left(\mathbf{x}_{\lfloor \frac{1}{\eta} \rfloor} - \mathbf{x}^*\right) = O(\eta).$$

2. Get the true gradient ∇F into the equation to form a martingale difference sequence

$$\gamma_\eta^{(k)} = \nabla F\left(\mathbf{x}_\eta^{(k)}\right) - \nabla f\left(\mathbf{x}_\eta^{(k)}, \xi_k\right)$$

How to get this SDE?

Basic technique [1, 2] to turn the stochastic algorithm around some local optimum $\mathbf{x}^*$ into some SDE

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right)$$

1. Work with error terms, i.e. normalize the variable by

$$\mathbf{u}_\eta^{(k)} = \frac{\mathbf{x}_\eta^{(k)} - \mathbf{x}^*}{\sqrt{\eta}}, \textit{ as } \mathrm{Var}\left(\mathbf{x}_{\lfloor \frac{1}{\eta} \rfloor} - \mathbf{x}^*\right) = O(\eta).$$

2. Get the true gradient ∇F into the equation to form a martingale difference sequence

$$\gamma_\eta^{(k)} = \nabla F\left(\mathbf{x}_\eta^{(k)}\right) - \nabla f\left(\mathbf{x}_\eta^{(k)}, \xi_k\right)$$

3. Identify the variance of all noises in the current setting.

How to get this SDE?

Basic technique [1, 2] to turn the stochastic algorithm around some local optimum $\mathbf{x}^*$ into some SDE

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - \eta \nabla f\left(\mathbf{x}^{(k-1)}; \xi_k\right)$$

1. Work with error terms, i.e. normalize the variable by

$$\mathbf{u}_\eta^{(k)} = \frac{\mathbf{x}_\eta^{(k)} - \mathbf{x}^*}{\sqrt{\eta}}, \textit{ as} \operatorname{Var}\left(\mathbf{x}_{\lfloor \frac{1}{\eta} \rfloor} - \mathbf{x}^*\right) = O(\eta).$$

2. Get the true gradient ∇F into the equation to form a martingale difference sequence

$$\gamma_\eta^{(k)} = \nabla F\left(\mathbf{x}_\eta^{(k)}\right) - \nabla f\left(\mathbf{x}_\eta^{(k)}, \xi_k\right)$$

3. Identify the variance of all noises in the current setting.
4. Take $\eta \to 0$ for the discretization scheme.

References

- Kushner, H. J. and Yin, G. G., Stochastic Approximation and Recursive Algorithms and Applications. Springer. 2003.
- Kushner, H. J., Stochastic Approximation: A Survey. 2008.
- Hu, W., Li, C. J., Li. L. and Liu J., On the diffusion approximation of nonconvex stochastic gradient descent. *Annals of Mathematical Sciences and Applications*. 2019.
- Glynn, P. M., Diffusion Approximations. Handbooks on OR & MS. 1990.
- Liu, T., Chen, Z., Zhou, E. and Zhao, T., A Diffusion Approximation Theory of Momentum SGD in Nonconvex Optimization.
- Komlós–Major–Tusnády approximation