A random perturbation approach to some stochastic approximation algorithms in optimization.

Wenqing Hu. ¹
(Presentation based on joint works with Chris Junchi Li², Weijie Su³, Haoyi Xiong⁴.)

¹ Department of Mathematics and Statistics, Missouri S&T.
² Department of Operations Research and Financial Engineering, Princeton University.
³ Department of Statistics, University of Pennsylvania.
⁴ Department of Computer Science, Missouri S&T.
Stochastic Optimization.

- **Stochastic Optimization Problem**
  
  $$
  \min_{x \in U \subset \mathbb{R}^d} F(x) .
  $$

- $$F(x) \equiv \mathbb{E}[F(x; \zeta)] .$$

- Index random variable $\zeta$ follows some prescribed distribution $\mathcal{D}$. 
Stochastic Gradient Descent Algorithm.

We target at finding a local minimum point $x^*$ of the expectation function $F(x) \equiv \mathbb{E}[F(x; \zeta)]$:

$$x^* = \arg \min_{x \in U \subset \mathbb{R}^d} \mathbb{E}[F(x; \zeta)].$$

We consider a general nonconvex stochastic loss function $F(x; \zeta)$ that is twice differentiable with respect to $x$. Together with some additional regularity assumptions\(^5\) we guarantee that $\nabla \mathbb{E}[F(x; \zeta)] = \mathbb{E}[\nabla F(x; \zeta)]$.

---

\(^5\) Such as control of the growth of the gradient in expectation.
Stochastic Gradient Descent Algorithm.

- Gradient Descent (GD) has iteration:
  \[ x(t) = x(t-1) - \eta \nabla F(x^{(t-1)}) , \]
  or in other words
  \[ x(t) = x(t-1) - \eta \mathbb{E}[\nabla F(x^{(t-1)}; \zeta)] . \]

- Stochastic Gradient Descent (SGD) has iteration:
  \[ x(t) = x(t-1) - \eta \nabla F(x^{(t-1)}; \zeta_t) , \]
  where \( \{ \zeta_t \} \) are i.i.d. random variables that have the same distribution as \( \zeta \sim D \).
Deep Neural Network (DNN). Goal is to solve the following stochastic optimization problem

$$\min_{x \in \mathbb{R}^d} F(x) \equiv \frac{1}{M} \sum_{i=1}^{M} F_i(x)$$

where each component $F_i$ corresponds to the loss function for data point $i \in \{1, \ldots, M\}$, and $x$ is the vector of weights being optimized.
Let $B$ be the minibatch of prescribed size uniformly sampled from $\{1, \ldots, M\}$, then the objective function can be further written as the expectation of a stochastic function

$$\frac{1}{M} \sum_{i=1}^{M} F_i(x) = E_B \left( \frac{1}{|B|} \sum_{x \in B} F_i(x) \right) \equiv E_B F(x; B).$$

SGD updates as

$$x(t) = x(t-1) - \eta \left( \frac{1}{|B_t|} \sum_{i \in B_t} \nabla F_i(x^{(t-1)}) \right),$$

which is the classical mini–batch version of the SGD.
Online learning via SGD: standard sequential predicting problem where for $i = 1, 2, ...$
1. An unlabeled example $a_i$ arrives;
2. We make a prediction $\hat{b}_i$ based on the current weights $x_i = [x_i^1, ..., x_i^d] \in \mathbb{R}^d$;
3. We observe $b_i$, let $\zeta_i = (a_i, b_i)$, and incur some known loss $F(x_i; \zeta_i)$ which is convex in $x_i$;
4. We update weights according to some rule $x_{i+1} \leftarrow S(x_i)$.

SGD rule:

$$S(x_i) = x_i - \eta \nabla_1 F(x_i; \zeta_i),$$

where $\nabla_1 F(u; \nu)$ is a subgradient of $F(u; \nu)$ with respect to the first variable $u$, and the parameter $\eta > 0$ is often referred as the learning rate.
Closer look at SGD.

- SGD:
  \[ x^{(t)} = x^{(t-1)} - \eta \nabla F(x^{(t-1)}; \zeta_t) \, . \]

- \( F(x) \equiv \mathbb{E} F(x; \zeta) \) in which \( \zeta \sim \mathcal{D} \).

- Let
  \[ e_t = \nabla F(x^{(t-1)}; \zeta_t) - \nabla F(x^{(t-1)}) \]

  and we can rewrite the SGD as
  \[ x^{(t)} = x^{(t-1)} - \eta (\nabla F(x^{(t-1)}) + e_t) \, . \]
Local statistical characteristics of SGD path.

- In a difference form it looks like

\[ x(t) - x(t-1) = -\eta \nabla F(x(t-1)) - \eta e_t. \]

- We see that

\[ \mathbb{E}(x(t) - x(t-1)|x(t-1)) = -\eta \nabla F(x(t-1)) \]

and

\[ \left[ \text{Cov}(x(t) - x(t-1)|x(t-1)) \right]^{1/2} = \eta \left[ \text{Cov}(\nabla F(x(t-1), \zeta)) \right]^{1/2}. \]
SGD approximating diffusion process.

- Very roughly speaking, we can approximate $x^{(t)}$ by a diffusion process $X_t$ driven by the stochastic differential equation

$$dX_t = -\eta \nabla F(X_t) dt + \eta \sigma(X_t) dW_t, \quad X_0 = x^{(0)},$$

where $\sigma(x) = [\text{Cov}(\nabla F(x; \zeta))]^{1/2}$ and $W_t$ is a standard Brownian motion in $\mathbb{R}^d$.

- Slogan: Continuous Markov processes are characterized by its local statistical characteristics only in the first and second moments (conditional mean and (co)variance).
Diffusion Approximation of SGD: Justification.

- Such an approximation has been justified in the weak sense in many classical literature\(^6\)\(^7\)\(^8\).
- It can also be thought of as a normal deviation result.
- One can call the continuous process \(X_t\) as the “continuous SGD”\(^9\).
- We also have a work in this direction (Hu–Li–Li–Liu, 2018).
- I will come back to this topic by the end of the talk.

---

SGD approximating diffusion process: convergence time.

- Recall that we have set $F(x) = \mathbb{E}F(x; \zeta)$.

- Thus we can formulate the original optimization problem as

$$x^* = \arg\min_{x \in U} F(x).$$

- Instead of SGD, in this work let us consider its approximating diffusion process

$$dX_t = -\eta \nabla F(X_t)dt + \eta \sigma(X_t) dW_t, \quad X_0 = x^{(0)}.$$

- The dynamics of $X_t$ is used as an alternative optimization procedure to find $x^*$.

- Remark: If there is no noise, then

$$dX_t = -\eta \nabla F(X_t)dt, \quad X_0 = x^{(0)}$$

is just the gradient flow $S^t x^{(0)}$. Thus it approaches $x^*$ in the strictly convex case.
SGD approximating diffusion process : convergence time.

- Hitting time

\[ \tau^{\eta} = \inf \{ t \geq 0 : F(X_t) \leq F(x^*) + \eta \} \]

for some small \( \eta > 0 \).

- Asymptotic of \( E\tau^{\eta} \) as \( \eta \to 0 \) ?
SGD approximating diffusion process: convergence time.

- Approximating diffusion process

\[ dX_t = -\eta \nabla F(X_t) dt + \eta \sigma(X_t) dW_t, \quad X_0 = x^{(0)}. \]

- Let \( Y_t = X_t/\eta \), then

\[ dY_t = -\nabla F(Y_t) dt + \sqrt{\eta} \sigma(Y_t) dW_t, \quad Y_0 = x^{(0)}. \]

(Random perturbations of the gradient flow!)

- Hitting time

\[ T^\eta = \inf\{ t \geq 0 : F(Y_t) \leq F(x^*) + e \} \]

for some small \( e > 0 \).

\[ \tau^\eta = \eta^{-1} T^\eta. \]
Random perturbations of the gradient flow: convergence time.

Let $Y_t = X_t/\eta$, then

$$dY_t = -\nabla F(Y_t)dt + \sqrt{\eta}\sigma(Y_t)dW_t , \ Y_0 = x^{(0)} .$$

Hitting time

$$T^\eta = \inf\{t \geq 0 : F(Y_t) \leq F(x^*) + e\}$$

for some small $e > 0$.

Asymptotic of $\mathbb{E}T^\eta$ as $\eta \to 0$?
Where is the difficulty?

**Figure 1:** Various critical points and the landscape of $F$. 
Where is the difficulty?

Figure 2: Higher order critical points.
Definition (strict saddle property)

Given fixed $\gamma_1 > 0$ and $\gamma_2 > 0$, we say a Morse function $F$ defined on $\mathbb{R}^n$ satisfies the “strict saddle property” if each point $x \in \mathbb{R}^n$ belongs to one of the following: (i) $|\nabla F(x)| \geq \gamma_2 > 0$; (ii) $|\nabla F(x)| < \gamma_2$ and $\lambda_{\min}(\nabla^2 F(x)) \leq -\gamma_1 < 0$; (iii) $|\nabla F(x)| < \gamma_2$ and $\lambda_{\min}(\nabla^2 F(x)) \geq \gamma_1 > 0$. 


11. Sun, J., Qu, Q., Wright, J., When are nonconvex problems not scary? arXiv:1510.06096[math.DC]
Strict saddle property.

**Figure 3**: Types of local landscape geometry with only strict saddles.
Strong saddle property.

- **Definition (strong saddle property)**

  Let the Morse function $F(\bullet)$ satisfy the strict saddle property with parameters $\gamma_1 > 0$ and $\gamma_2 > 0$. We say the Morse function $F(\bullet)$ satisfy the “strong saddle property” if for some $\gamma_3 > 0$ and any $x \in \mathbb{R}^n$ such that $\nabla F(x) = 0$, all eigenvalues $\lambda_i$, $i = 1, 2, \ldots, n$ of the Hessian $\nabla^2 F(x)$ at $x$ satisfying (ii) in Definition 1 are bounded away from zero by some $\gamma_3 > 0$ in absolute value, i.e., $|\lambda_i| \geq \gamma_3 > 0$ for any $1 \leq i \leq n$. 
Escape from saddle points.

**Figure 4:** Escape from a strong saddle point.
Escape from saddle points: behavior of process near one specific saddle.

- The problem was first studied by Kifer in 1981\(^{12}\).
- Recall that

\[
dY_t = -\nabla F(Y_t) dt + \sqrt{\eta} \sigma(Y_t) dW_t , \quad Y_0 = x^{(0)},
\]

is a random perturbation of the gradient flow. Kifer needs to assume further that the diffusion matrix

\[
\sigma(\bullet)\sigma^T(\bullet) = \text{Cov}(\nabla F(\bullet; \zeta))
\]

is strictly uniformly positive definite.

- Roughly speaking, Kifer’s result states that the exit from a neighborhood of a strong saddle point happens along the “most unstable” direction for the Hessian matrix, and the exit time is asymptotically \( \sim C \log(\eta^{-1}) \)

Escape from saddle points: behavior of process near one specific saddle.

**Figure 5:** Escape from a strong saddle point: Kifer's result.
Escape from saddle points: behavior of process near one specific saddle.

- $G \cup \partial G = 0 \cup A_1 \cup A_2 \cup A_3$.
- If $x \in A_2 \cup A_3$, then for the deterministic gradient flow there is a finite
  \[ t(x) = \inf\{ t > 0 : S^t x \in \partial G \} \, . \]
- In this case as $\eta \to 0$, expected first exit time converges to $t(x)$, and first exit position converges to $S^{t(x)} x$.
- Remind that the gradient flow $S^t x^{(0)}$ is
  \[ dX_t = -\eta \nabla F(X_t) dt \, , \, X_0 = x^{(0)} \, . \]
Escape from saddle points: behavior of process near one specific saddle.

**Figure 6:** Escape from a strong saddle point: Kifer's result.
Escape from saddle points: behavior of process near one specific saddle.

- If $x \in (0 \cup A_1) \setminus \partial G$, situation is more interesting.
- First exit occurs along the direction pointed out by the most negative eigenvalue(s) of the Hessian $\nabla^2 F(0)$.
- $\nabla^2 F(0)$ has spectrum $-\lambda_1 = -\lambda_2 = ... = -\lambda_q < -\lambda_{q+1} \leq ... \leq -\lambda_p < 0 < \lambda_{p+1} \leq ... \leq \lambda_d$.
- Exit time will be asymptotically $\frac{1}{2\lambda_1} \ln(\eta^{-1})$ as $\eta \to 0$.
- Exit position will converge (in probability) to the intersection of $\partial G$ with the invariant manifold corresponding to $-\lambda_1, ..., -\lambda_q$.
- Slogan: when the process $Y_t$ comes close to 0, it will “choose” the most unstable directions and move along them.
Escape from saddle points: behavior of process near one specific saddle.

Figure 7: Escape from a strong saddle point: Kifer’s result.
More technical aspects of Kifer’s result...

- In Kifer’s result all convergence are point–wise with respect to initial point $x$. They are not uniform convergence when $\eta \to 0$ with respect to all initial point $x$.
- The “exit along most unstable directions” happens only when initial point $x$ strictly stands on $A_1$.
- Does not allow small perturbations with respect to initial point $x$. Run into messy calculations...
Why these technicalities matter?

- Our landscape may have a \textit{“chain of saddles”}.
- Recall that

\[ dY_t = -\nabla F(Y_t) dt + \sqrt{\eta} \sigma(Y_t) dW_t, \quad Y_0 = x^{(0)} \]

is a random perturbation of the gradient flow.
Global landscape : chain of saddle points.

Figure 8: Chain of saddle points.
Linerization of the gradient flow near a strong saddle point.

- **Hartman–Grobman Theorem**: for any strong saddle point $O$ that we consider, there exist an open neighborhood $U$ of $O$ and a $C^0$ homeomorphism $h : U \to \mathbb{R}^d$ such that the gradient flow under $h$ is mapped into a linear flow.

- **Linearization Assumption**: The homeomorphism $h$ provided by the Hartman–Grobman theorem can be taken to be $C^2$.

- A sufficient condition for the validity of the $C^2$ linearization assumption is the so-called *non-resonance condition* (Sternberg linearization theorem).

- I will also come back to this topic at the end of the talk.

---

Linerization of the gradient flow near a strong saddle point.

**Figure 9:** Linerization of the gradient flow near a strong saddle point.
Uniform version of Kifer’s exit asymptotic.

- Let $U \subset G$ be an open neighborhood of the saddle point $O$. Set initial point $x \in U \cup \partial U$.
- Set $\text{dist}(U \cup \partial U, \partial G) > 0$. Let $t(x) = \inf\{t > 0 : S^tx \in \partial G\}$.
- $W_{\text{max}}$ is the invariant manifold corresponding to the most negative eigenvalues $-\lambda_1, ..., -\lambda_q$ of $\nabla^2 F(O)$.
- Define $Q_{\text{max}} = W_{\text{max}} \cap \partial G$.
- Set
  
  $$\partial G_{U \cup \partial U \to \text{out}} = \{S^t(x)x \text{ for some } x \in U \cup \partial U \text{ with finite } t(x)\} \cup Q_{\text{max}}$$

- For small $\mu > 0$ let
  
  $$Q^\mu = \{x \in \partial G, \text{dist}(x, \partial G_{U \cup \partial U \to \text{out}}) < \mu\}.$$ 

- $\tau_x^n$ is the first exit time to $\partial G$ for the process $Y_t$ starting from $x$. 
Uniform version of Kifer’s exit asymptotic.

\[ D_i(F(O_i) + \frac{h}{2}, F(O_i-1) - \frac{h}{2}) \]

\[ C_i(F(O_i) + h) \]

\[ C_i(F(O_i) + \frac{h}{2}) \]

\[ C_i(F(O_i) - h) \]

\( \{ x \in \partial V_i : \text{dist}(x, \partial V_i, \bar{U}_i \rightarrow \text{out}) < \mu \} \)

**Figure 10:** Uniform exit dynamics near a specific strong saddle point under the linerization assumption.
Uniform version of Kifer’s exit asymptotic.

Theorem

For any $r > 0$, there exist some $\eta_0 > 0$ so that for all $x \in U \cup \partial U$ and all $0 < \eta < \eta_0$ we have

$$\frac{E_x \tau_x^\eta}{\ln(\eta^{-1})} \leq \frac{1}{2\lambda_1} + r.$$  

For any small $\mu > 0$ and any $\rho > 0$, there exist some $\eta_0 > 0$ so that for all $x \in U \cup \partial U$ and all $0 < \eta < \eta_0$ we have

$$P_x( Y_{\tau^\eta_x} \in Q^\mu ) \geq 1 - \rho.$$
Convergence analysis in a basin containing a local minimum: Sequence of stopping times.

To demonstrate our analysis, we will assume that $x^*$ is a local minimum point of $F(\bullet)$ in the sense that for some open neighborhood $U(x^*)$ of $x^*$ we have

$$x^* = \arg\min_{x \in U(x^*)} F(x).$$

Assume that there are $k$ strong saddle points $O_1, ..., O_k$ in $U(x^*)$ such that $F(O_1) > F(O_2) > ... > F(O_k) > F(x^*)$.

Start the process with initial point $Y_0 = x \in U(x^*)$.

Hitting time

$$T^n = \inf\{ t \geq 0 : F(Y_t) \leq F(x^*) + e \}$$

for some small $e > 0$.

Asymptotic of $E T^n$ as $\eta \to 0$?
Convergence analysis in a basin containing a local minimum: Sequence of stopping times.

**Figure 11:** Sequence of stopping times.
Convergence analysis in a basin containing a local minimum: Sequence of stopping times.

- Standard Markov cycle type argument.
- But the geometry is a little different from the classical arguments for elliptic equilibriums found in Freidlin–Wentzell book\textsuperscript{14}.

\[ \mathbb{E} T^n \lesssim \frac{k}{2\gamma_1} \ln(\eta^{-1}) \] conditioned upon convergence.

Recall that we have set $F(x) = \mathbb{E} F(x; \zeta)$.

Thus we can formulate the original optimization problem as

$$x^* = \arg\min_{x \in U(x^*)} F(x).$$

Instead of SGD, in this work let us consider its approximating diffusion process

$$dX_t = -\eta \nabla F(X_t) dt + \eta \sigma(X_t) dW_t, \quad X_0 = x^{(0)}.$$

The dynamics of $X_t$ is used as an surrogate optimization procedure to find $x^*$.
SGD approximating diffusion process: convergence time.

- Hitting time

\[ \tau^n = \inf\{t \geq 0 : F(X_t) \leq F(x^*) + e\} \]

for some small \( e > 0 \).

- Asymptotic of \( E\tau^n \) as \( \eta \to 0 \) ?
SGD approximating diffusion process: convergence time.

- Approximating diffusion process

\[ dX_t = -\eta \nabla F(X_t) \, dt + \eta \sigma(X_t) \, dW_t, \quad X_0 = x^{(0)}. \]

- Let \( Y_t = X_t/\eta \), then

\[ dY_t = -\nabla F(Y_t) \, dt + \sqrt{\eta} \sigma(Y_t) \, dW_t, \quad Y_0 = x^{(0)}. \]

(Random perturbations of the gradient flow!)

- Hitting time

\[ T^\eta = \inf \{ t \geq 0 : F(Y_t) \leq F(x^*) + e \} \]

for some small \( e > 0 \).

\[ \tau^\eta = \eta^{-1} T^\eta. \]
Convergence analysis in a basin containing a local minimum.

- Theorem
  (i) For any small $\rho > 0$, with probability at least $1 - \rho$, SGD approximating diffusion process $X_t$ converges to the minimizer $x^*$ for sufficiently small $\beta$ after passing through all $k$ saddle points $O_1$, $\ldots$, $O_k$;
  (ii) Consider the stopping time $\tau^\eta$. Then as $\eta \downarrow 0$, conditioned on the above convergence of SGD approximating diffusion process $X_t$, we have

$$
\lim_{\eta \to 0} \frac{\mathbf{E}_{\tau^\eta}}{\eta^{-1} \ln \eta^{-1}} \leq \frac{k}{2\gamma_1}.
$$
Philosophy.

discrete stochastic optimization algorithms
\Downarrow

diffusion approximation

\leftarrow \text{analysis of convergence time}
\Uparrow

\leftarrow \text{random perturbations of dynamical systems}
Other Examples: Stochastic heavy–ball method.

\[
\min_{x \in \mathbb{R}^d} f(x) .
\]

”heavy ball method” (B.Polyak 1964)

\[
\begin{cases}
  x_t = x_{t-1} + \sqrt{s}v_t; \\
  v_t = (1 - \alpha \sqrt{s})v_{t-1} - \sqrt{s}\nabla x f (x_{t-1} + \sqrt{s}(1 - \alpha \sqrt{s})v_{t-1}); \\
  x_0, v_0 \in \mathbb{R}^d .
\end{cases}
\]

Adding momentum leads to the second–order equation

\[
\ddot{X}(t) + \alpha \dot{X}(t) + \nabla X f(X(t)) = 0 , \ X(0) \in \mathbb{R}^d .
\]

Hamiltonian system for a dissipative nonlinear oscillator:

\[
\begin{cases}
  dX(t) = V(t)dt , \ X(0) \in \mathbb{R}^d ; \\
  dV(t) = (-\alpha V(t) - \nabla X f(X(t)))dt , \ V(0) \in \mathbb{R}^d .
\end{cases}
\]
Other Examples: Stochastic heavy–ball method.

- How to escape from saddle points of $f(x)$?
- Add noise → Randomly perturbed Hamiltonian system for a dissipative nonlinear oscillator:

$$\begin{align*}
  dX^\varepsilon_t &= V^\varepsilon_t \, dt + \varepsilon \sigma_1(X^\varepsilon_t, V^\varepsilon_t) \, dW^1_t; \\
  dV^\varepsilon_t &= (-\alpha V^\varepsilon_t - \nabla_x f(X^\varepsilon_t)) \, dt + \varepsilon \sigma_2(X^\varepsilon_t, V^\varepsilon_t) \, dW^2_t, \\
  X^\varepsilon_0, V^\varepsilon_0 &\in \mathbb{R}^d.
\end{align*}$$
Diffusion limit of stochastic heavy–ball method: Randomly perturbed Hamiltonian system.

Figure 12: Randomly perturbed Hamiltonian system.
Diffusion limit of stochastic heavy–ball method: Randomly perturbed Hamiltonian system.

**Figure 13:** Phase Diagram of a Hamiltonian System.
Diffusion limit of stochastic heavy–ball method: Randomly perturbed Hamiltonian system.

**Figure 14:** Behavior near a saddle point.
Diffusion limit of stochastic heavy–ball method: Randomly perturbed Hamiltonian system.

**Figure 15:** Chain of saddle points in a randomly perturbed Hamiltonian system.
Stochastic Composite Gradient Descent.

- By introducing approximating diffusion processes, method of stochastic dynamical systems, in particular random perturbations of dynamical systems are effectively used to analyze these limiting processes.
- For example, averaging principle can be used to analyze Stochastic Composite Gradient Descent (SCGD) (Hu–Li, 2017). We skip details here.
Philosophy.

- discrete stochastic optimization algorithms
- diffusion approximation

\[ \downarrow \]

\[ \uparrow \]

- analysis of convergence time
- random perturbations of dynamical systems
Inspiring practical algorithm design.

- The above philosophy of “random perturbations” also inspires practical algorithm design.
- What if the objective function $f(x)$ is even unknown, but needs to be learned?
- Work in progress with Xiong, H. from Missouri S&T Computer Science.
Epilogue : A few remaining problems.

- Approximating diffusion process $X_t$ : how it approximates $x(t)$? Various ways\textsuperscript{15}. May need correction term\textsuperscript{16} from numerical SDE theory.

- Linerization assumption : typical in dynamical systems, but shall be removed by much harder work.

- Attacking the discrete iteration \textit{directly} : Only works for special cases such as Principle Component Analysis (PCA)\textsuperscript{17}.

\textsuperscript{15} Li, Q., Tai, C., E. W., Stochastic modified equations and adaptive stochastic gradient algorithms, arXiv:1511.06251v3


\textsuperscript{17} Li, C.J., Wang, M., Liu, H., Zhang, T., Near–Optimal Stochastic Approximation for Online Principal Component Estimation, \textit{Mathematical Programming}, \textbf{167}(1), pp. 75–97, 2018
Thank you for your attention!