New Analysis of Adaptive Stochastic Optimization Methods via Martingales.

Katya Scheinberg

joint work with J. Blanchet (Stanford), C. Cartis (Oxford), M. Menickelly (Argonne) and C. Paquette (Waterloo)

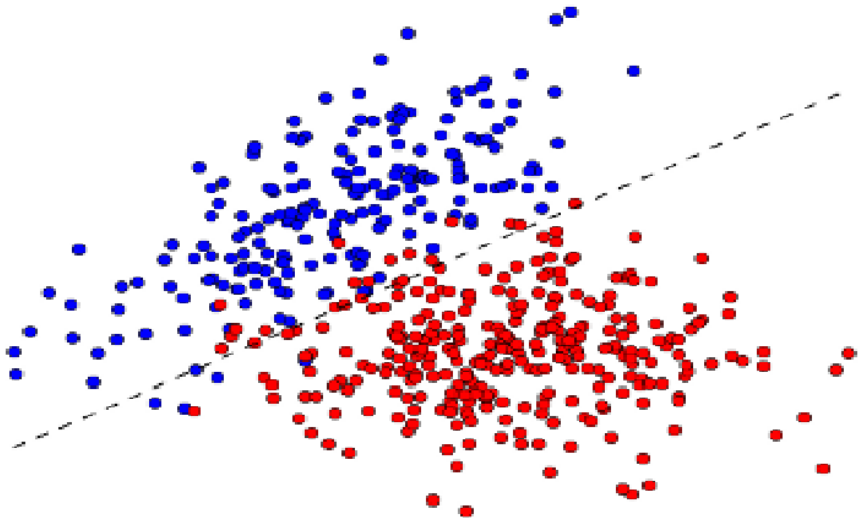
Princeton Day of Optimization

September 28, 2018

Outline

Outline

Introduction: Supervised Learning Problem



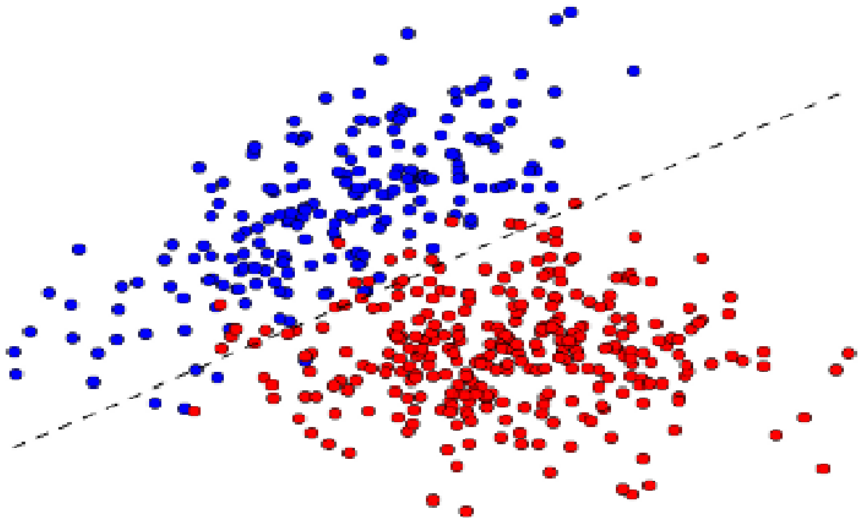
Example: binary classification

- Map $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ to $y \in \mathcal{Y} \subseteq \{0, 1\}$.
- Consider predictors in the form $p(x; w)$ so that

$$p(\cdot, w) : \mathcal{X} \to \mathcal{Y},$$

- If $p(x, w) = w^T x$ - linear classifier, more generally $p(x, w)$ is nonlinear, e.g., neural network.

Introduction: Supervised Learning Problem



How do we select the best classifier?

- Min Expected/Empirical Error
- Min Expected/Empirical Loss
- Max Expected/Empirical AUC

Example: binary classification

- Map $x \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ to $y \in \mathcal{Y} \subseteq \{0, 1\}$.
- Consider predictors in the form $p(x; w)$ so that

$$p(\cdot, w) : \mathcal{X} \to \mathcal{Y},$$

- If $p(x, w) = w^T x$ - linear classifier, more generally $p(x, w)$ is nonlinear, e.g., neural network.

How to find the best predictor?

$$\min_{w \in \mathcal{W}} \; f(w) := \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}[yp(w,x) \leq 0] dP(x,y).$$

"Intractable" because of unknown distribution.

- Instead use empirical risk of $p(x;w)$ over the finite training set $\mathcal{S}$,

$$\min_{w \in \mathcal{W}} \; f_{\mathcal{S}}(w) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[y_i p(w,x_i) \leq 0].$$

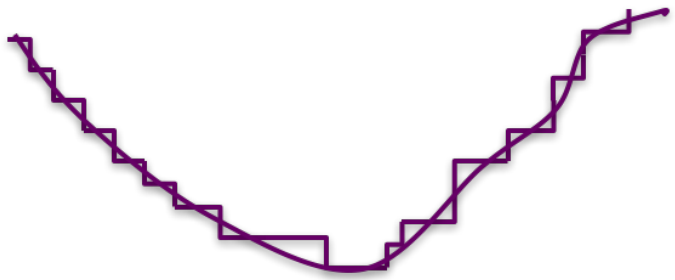
How to find the best predictor?

$$\min_{w \in \mathcal{W}} \; f(w) := \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}[yp(w,x) \leq 0] dP(x,y).$$

"Intractable" because of unknown distribution.

- Instead use empirical risk of $p(x;w)$ over the finite training set $\mathcal{S}$,

$$\min_{w \in \mathcal{W}} \; f_{\mathcal{S}}(w) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[y_i p(w,x_i) \leq 0].$$



Hard problem?

How to find the best predictor?

$$\min_{w \in \mathcal{W}} \; f(w) := \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{1}[yp(w,x) \leq 0] dP(x,y).$$

"Intractable" because of unknown distribution.

- Instead use empirical risk of $p(x;w)$ over the finite training set $\mathcal{S}$,

$$\min_{w \in \mathcal{W}} \; f_{\mathcal{S}}(w) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[y_i p(w,x_i) \leq 0].$$

- Instead use "smooth" or "easy" empirical loss $p(x;w)$ over the finite training set $\mathcal{S}$,

$$\min_{w \in \mathcal{W}} \; \hat{f}_{\ell}(w) := \frac{1}{n} \sum_{i=1}^{n} \ell(p(w,x_i), y_i).$$

This is a tractable problem but n and d_w can be large!!

What's been happening in optimization under the influence of machine learning applications?

- "New" scale - optimizing very large sums
- Stochasticity - optimizing averages and/or expectations
- Inexactness - optimizing using "cheap" inexact steps
- Complexity - emphasis on complexity bounds

Stochastic Optimization Setting

$$\text{Minimize } f(x) : \mathbb{R}^n \to \mathbb{R}$$

- We will assume throughout that f is L-smooth and bounded below.
- f is nonconvex unless specified.
- $f(x)$ is stochastic, i.e. given x, we obtain estimates $\tilde{f}(x, \xi)$, (and possibly $\nabla \tilde{f}(x, \xi)$), where ξ is a random variable.
- Note: $f(x)$ and $\nabla f(x)$ are not necessarily equal to $\mathbb{E}_\xi[\tilde{f}(x, \xi)]$, $\mathbb{E}_\xi[\nabla \tilde{f}(x, \xi)]$

Motivation #1: Adaptive Methods

Standard stochastic gradient method

- Assume $\mathbb{E}_\xi[\nabla \tilde{f}(x^k, \xi)] = \nabla f(x^k)$ for all x^k.
- Generate $\xi_1, \xi_2, \ldots, \xi_{s^k}$ using a predefined sample (mini-batch) size s^k and compute

$$\nabla_{s^k} f(x) = \frac{1}{s^k} \sum_{i=1}^{s^k} \nabla \tilde{f}(x^k, \xi_i).$$

- Take a step $x^{k+1} = x^k - \alpha_k \nabla_{s^k} f(x)$ with predefined step size α_k (learning rate).

Motivation #1: Adaptive Methods

Standard stochastic gradient method

- Assume $\mathbb{E}_\xi[\nabla \tilde{f}(x^k, \xi)] = \nabla f(x^k)$ for all x^k.
- Generate $\xi_1, \xi_2, \ldots, \xi_{s^k}$ using a predefined sample (mini-batch) size s^k and compute

$$\nabla_{s^k} f(x) = \frac{1}{s^k} \sum_{i=1}^{s^k} \nabla \tilde{f}(x^k, \xi_i).$$

- Take a step $x^{k+1} = x^k - \alpha_k \nabla_{s^k} f(x)$ with predefined step size α_k (learning rate).

In deterministic optimization adaptive methods are more efficient!

Motivation #1: Adaptive Methods

Standard stochastic gradient method

- Assume $\mathbb{E}_\xi[\nabla \tilde{f}(x^k, \xi)] = \nabla f(x^k)$ for all x^k.
- Generate $\xi_1, \xi_2, \ldots, \xi_{s^k}$ using a predefined sample (mini-batch) size s^k and compute

$$\nabla_{s^k} f(x) = \frac{1}{s^k} \sum_{i=1}^{s^k} \nabla \tilde{f}(x^k, \xi_i).$$

- Take a step $x^{k+1} = x^k - \alpha_k \nabla_{s^k} f(x)$ with predefined step size α_k (learning rate).

In deterministic optimization adaptive methods are more efficient!

Can we design and analyze adaptive stochastic methods?

(Deterministic) Backtracking Line Search

Backtracking Line Search Algorithm

- Choose initial point x^0, initial step-size $\alpha_0 \in (0, \alpha_{\max})$ with $\alpha_{\max} > 0$. Choose constant $\gamma > 1$, set $k \leftarrow 0$.
- For $k = 0, 1, 2, \ldots$
- Compute a descent direction $-g_k$.
- Compute $f(x_k)$ and $f(x_k - \alpha_k g_k)$ and check sufficient decrease:
$$f(x_k - \alpha_k g_k) \leq f(x_k) - \theta \alpha_k \|g_k\|^2.$$
- Successful: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ and $\alpha_{k+1} = \min\{\gamma \alpha_k, \alpha_{max}\}$.
- Unsuccessful: $x_{k+1} = x_k$ and $\alpha_{k+1} = \alpha_k / \gamma$.

(Deterministic) Trust Region Method

- Choose initial point x^0, initial trust-region radius $\delta_0 \in (0, \delta_{\max})$ with $\delta_{\max} > 0$. Choose constants $\gamma > 1$, $\eta_1 \in (0, 1)$, $\eta_2 > 0$. Set $k \leftarrow 0$.
- For $k = 0, 1, 2, \ldots$
- Build a local model $m_k(\cdot)$ of f on $\mathcal{B}(0, \delta_k)$.
- Compute $s^k = \arg \min_{s:\|s\| \le \delta_k} m_k(s)$ (approximately).
- Compute $\rho_k = \dfrac{f(x^k) - f(x^k + s^k)}{m_k(x^k) - m_k(x^k + s^k)}$.
- If $\rho_k \ge \eta_1$ and $\|\nabla m_k(x^k)\| \ge \eta_2 \delta_k$ then $x^{k+1} = x^k + s^k$ (Successful) else $x^{k+1} = x^k$ (Unsuccessful).
- If successful then $\delta_{k+1} = \min\{\gamma \delta_k, \delta_{\max}\}$
  If unsuccessful then $\delta_{k+1} = \gamma^{-1} \delta_k$.

Complexity Bounds

Deterministic Line Search and Trust Region

	$\|\nabla f(x_k)\| < \varepsilon$	$f(x_k) - f^* < \varepsilon$
L-smooth	$\frac{L}{\varepsilon^2}$	-
L-smooth/convex	$\frac{L}{\varepsilon^2}$	$\frac{L}{\varepsilon}$
μ-strongly convex	$\frac{L}{\mu} \cdot \log(\frac{1}{\varepsilon})$	$\frac{L}{\mu} \cdot \log(\frac{1}{\varepsilon})$

Stochastic Gradient

	$\|\nabla f(x_k)\| < \varepsilon$	$f(x_k) - f^* < \varepsilon$
L-smooth	$\frac{L}{\varepsilon^4}$	-
L-smooth/convex	$\frac{L}{\varepsilon^4}$	$\frac{L}{\varepsilon^2}$
μ-strongly convex	$\frac{L}{\varepsilon^2}$	$\frac{L}{\varepsilon}$

Motivation #2: Recovering Deterministic Complexity

Various adaptive stochastic methods that have partial convergence rates results matching deterministic rates.

- Friedlander and Schmidt. "Hybrid deterministic-stochastic methods for data fitting". *SIAM Journal on Scientific Computing* 34(3), 2012.
  Convergence rate for strongly convex case based on exponentially increasing sample size.
- Hashemi, Ghosh, and Pasupathy. "On adaptive sampling rules for stochastic recursions". *Simulation Conference (WSC), IEEE*, 2014.
  No rates, convergence with probability one by increasing sample size indefinitely.
- Bollapragada, Byrd, and Nocedal. "Adaptive sampling strategies for stochastic optimization". *arXiv:1710.11258*, 2017.
  Good heuristic ideas, convergence rates only when sample size chosen based on the true gradient.

Motivation #2: Recovering Deterministic Complexity

Various adaptive stochastic methods that have partial convergence rates results matching deterministic rates.

- Roosta-Khorasani and Mahoney. "Sub-sampled newton methods I: Globally convergent algorithms". *arXiv:1601.04737*, 2017
  Only Hessian estimates are stochastic, fixed large sample size, complexity bound established only under assumption that estimates are accurate at each iteration.
- Tripuraneni, Stern, Jin, Regier, and Jordan. "Stochastic cubic regularization for fast nonconvex optimization". *arXiv:1711.02838*, 2017
  Fixed large sample sizes, fixed small step size, complexity bound established only under assumption that estimates are accurate at each iteration.

Full convergence rates results are needed

Motivation #3: Expected Convergence Rate vs. Expected Complexity

- Complexity bound in deterministic optimization is the bound on the first iteration T_ϵ that achieves ϵ-accuracy, e.g. for nonconvex f

$$\|\nabla f(x^k)\| \leq \epsilon \Rightarrow T_\epsilon \leq O(\frac{1}{\epsilon^2})$$

  Then convergence rate is

$$\|\nabla f(x^k)\|^2 \leq O(\frac{1}{T})$$

- Convergence rate for stochastic algorithms for nonconvex f is the expected accuracy at the output $x^{\mathcal{T}}$ (usually random) produced after first T iterations

$$\mathbb{E}[\|\nabla f(x^{\mathcal{T}})\|^2] \leq \frac{1}{\sqrt{T}}$$

- In some cases bound on T_ϵ is only derived with high probability.

Can we bound $\mathbb{E}[T_\epsilon]$?

Prior works that bound $\mathbb{E}[T_\epsilon]$

- Cartis and Scheinberg. "Global convergence rate analysis of unconstrained optimization methods based on probabilistic models". *Mathematical Programming* 2017.
  First-order rates for line-search in nonconvex, convex and strongly convex cases, adaptive cubic regularization for nonconvex case.
- Gratton, Royer, Vicente, and Zhang. "Complexity and global rates of trust-region methods based on probabilistic models", *IMA Journal of Numerical Analysis*, 2018.
  First- and second-order rates for trust region in the nonconvex case.

Only Hessian and gradient estimates are stochastic

Motivation #4: Biased Estimates

Consider example:

$$f(x) = \sum_{i=1}^{n} (x_i - 1)^2$$

$$\tilde{f}(x, \xi) = \sum_{i=1}^{n} w_i, \text{ where } w_i = \begin{cases} (x^i - 1)^2 & \text{w.p. } p \\ -10000 & \text{w.p. } 1 - p \end{cases}$$

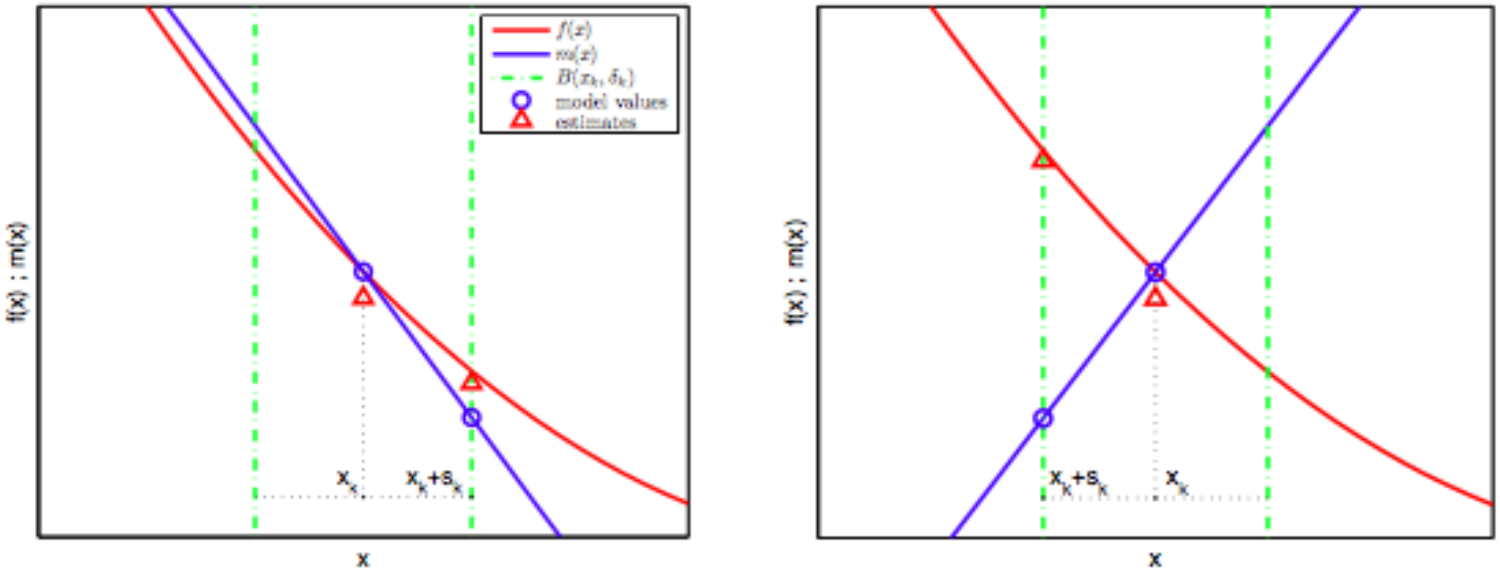
- For $p < 1$, this noise is biased ($\mathbb{E}_\xi[\tilde{f}(x, \xi)] \neq f(x)$).
- Does not fit into the assumptions of standard algorithms and analysis.
- Can we have convergence if p is sufficiently large?

Outline

Stochastic Trust Region Method

- Choose initial point x^0, initial trust-region radius $\delta_0 \in (0, \delta_{\max})$ with $\delta_{\max} > 0$. Choose constants $\gamma > 1$, $\eta_1 \in (0, 1)$, $\eta_2 > 0$. Set $k \leftarrow 0$.
- For $k = 0, 1, 2, \ldots$
- Build a local random model $m_k(\cdot)$ of f on $\mathcal{B}(0, \delta_k)$.
- Compute $s^k = \arg \min_{s:\|s\| \leq \delta_k} m_k(s)$ (approximately).
- Compute $\rho_k = \dfrac{f(x^k) - f(x^k + s^k)}{m_k(x^k) - m_k(x^k + s^k)}$.
- If $\rho_k \geq \eta_1$ and $\|\nabla m_k(x^k)\| \geq \eta_2 \delta_k$ then $x^{k+1} = x^k + s^k$ (Successful) else $x^{k+1} = x^k$ (Unsuccessful).
- If successful then $\delta_k \uparrow$
  If unsuccessful then $\delta_k \downarrow$

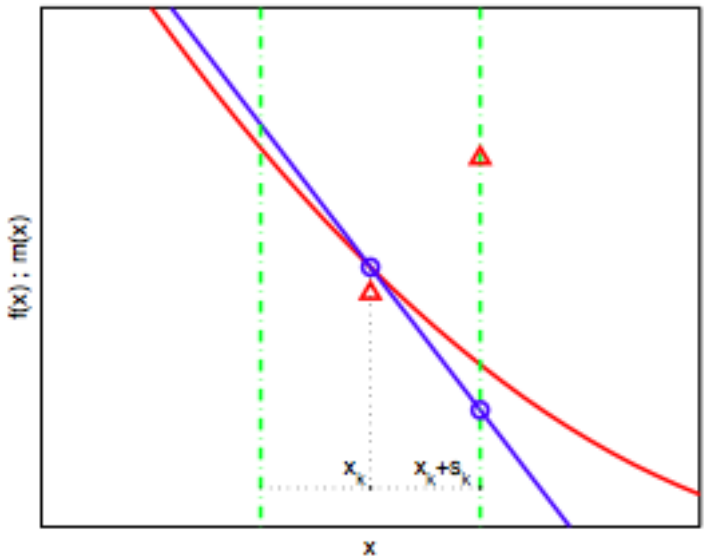
What can happen?



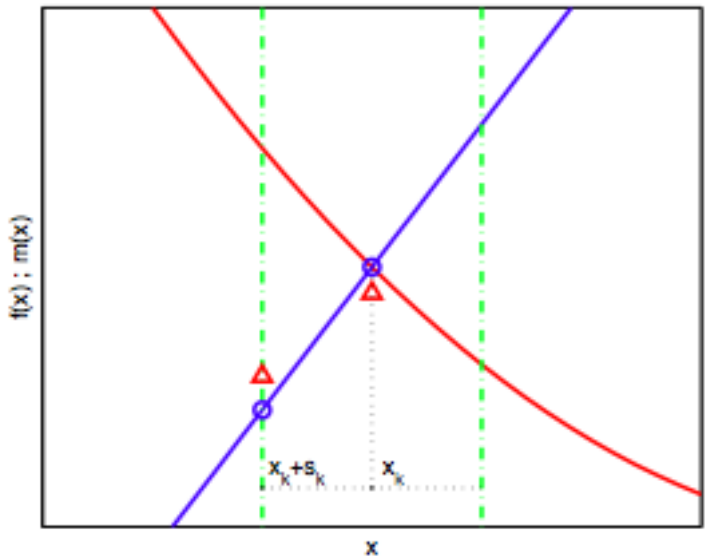
(a) Good model; good estimates.
True successful steps.

(b) Bad model; good estimates.
Unsuccessful steps.

What else can happen



(c) Good model; bad estimates.
Unsuccessful steps.



(d) Bad model; bad estimates.
False successful steps: f can increase!

Stochastic Line Search

- Compute random estimate of the gradient, g_k
- Compute random estimates $f_k^0 \approx f(x_k)$ and $f_k^s \approx f(x_k - \alpha_k g_k)$
- Check the sufficient decrease

$$f_k^s \leq f_k^0 - \theta \alpha_k \|g_k\|^2$$

- Successful: $x_{k+1} = x_k - \alpha_k g_k$ and $\alpha_k \uparrow$

  - Reliable step: If $\alpha_k \|g_k\|^2 \geq \delta_k^2$, $\quad \delta_k \uparrow$
  - Unreliable step: If $\alpha_k \|g_k\|^2 < \delta_k^2$, $\quad \delta_k \downarrow$

- Unsucessful: $x_{k+1} = x_k$, $\alpha_k \downarrow$, and $\boxed{\delta_k \downarrow}$

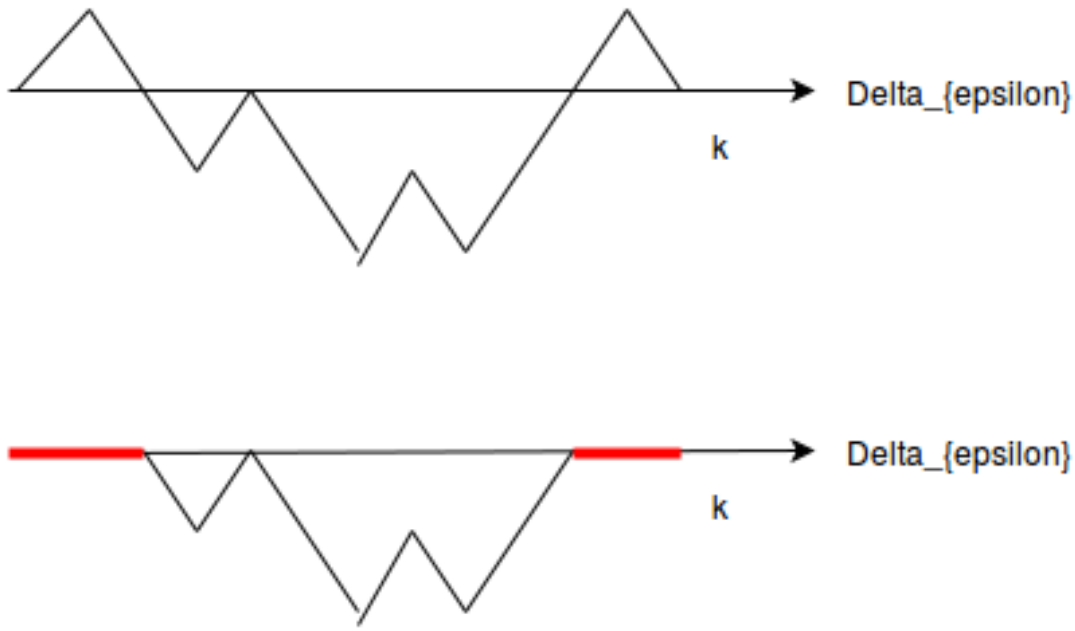
Outline

Stochastic Process (Algorithm)

Consider a stochastic process $\{\Phi_k, \Delta_k\} \geq 0$ and its stopping time T_ϵ.
Intuitively:

- Φ_k is progress towards optimality (order of $f(x^k) - f_{min}$).
- Δ_k is the step size parameter (trust region radius/learning rate).
- T_ϵ is the first iteration k to reach accuracy ϵ ($\|\nabla f(X^k)\| \leq \epsilon$).

The Δ_k Process



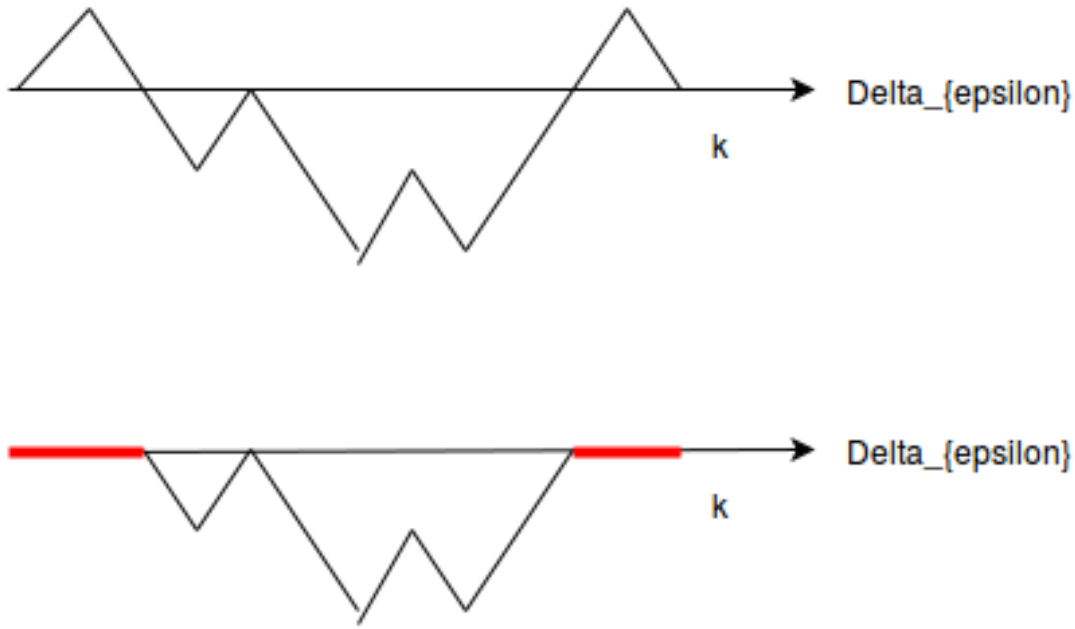
Key Assumption (i):

There exists a constant Δ_ϵ such that, until the stopping time:

$$\Delta_{k+1} \geq \min(\gamma^{W_{k+1}}\Delta_k, \Delta_\epsilon),$$

where W_{k+1} is a random walk with positive drift ($p > \frac{1}{2}$).

The Δ_k Process

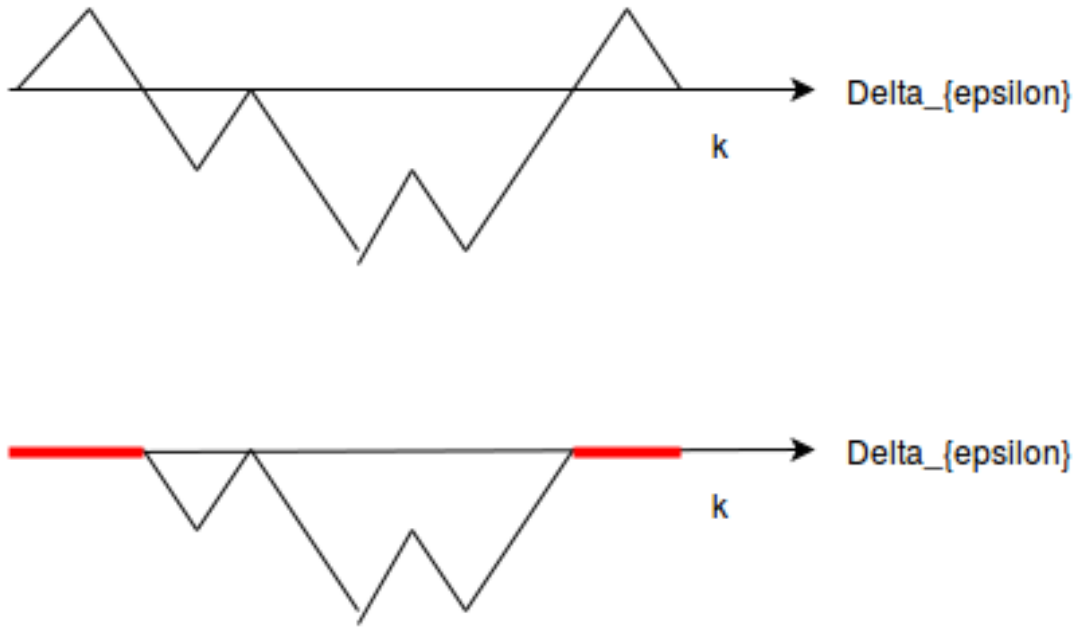


Key Assumption (ii):

There exists a nondecreasing function $h(\cdot) : [0, \infty) \to (0, \infty)$ and a constant $\Theta > 0$ such that, until the stopping time:

$$\mathbb{E}(\Phi_{k+1}|\mathcal{F}_k) \leq \Phi_k - \Theta h(\Delta_k).$$

The Δ_k Process



Main Idea:

- Δ_k may get arbitrarily small, but it tends to return to Δ_ϵ.
- Δ_k is a birth-death process with known interarrival times depending p.
- Count the returns of Δ_k to Δ_ϵ - a *renewal cycle*.
- For each renewal the is a fixed expected reward.

Bounding expected stopping time

Main Idea: This is a renewal-reward process and Φ_k is a supermartingale and $\Phi_0 \geq \Theta \sum_{i=0}^{T_\epsilon} h(\Delta_i)$.

- T_ϵ is a stopping time!
- Applying Wald's Identity we can bound the number of renewals that will occur before T_ϵ.
- Multiply by the expected renewal time.

We have the following results

Theorem (Blanchet, Cartis, Menickelly, S. '17)

Let the Key Assumption hold. Then

$$\mathbb{E}[T_\epsilon] \leq \frac{p}{2p-1} \cdot \frac{\Phi_0}{\Theta h(\Delta_\epsilon)} + 1.$$

Outline

Assumptions on models and estimates

For trust region, first-order

$$\begin{aligned} \|\nabla f(x^k) - \nabla m_k(x^k)\| &\leq \mathcal{O}(\delta_k), \quad \text{w.p. } p_g \\ |f_k^0 - f(x^k)| \leq \mathcal{O}(\delta_k^2) \text{ and } |f_k^s - f(x^k + s^k)| &\leq \mathcal{O}(\delta_k^2). \quad \text{w.p. } p_f \end{aligned}$$

For trust region, second-order

$$\begin{aligned} \|\nabla^2 f(x^k) - \nabla^2 m_k(x^k)\| &\leq \mathcal{O}(\delta_k) \\ \|\nabla f(x^k) - \nabla m_k(x^k)\| &\leq \mathcal{O}(\delta_k^2), \quad \text{w.p. } p_g \\ |f_k^0 - f(x^k)| \leq \mathcal{O}(\delta_k^2) \text{ and } |f_k^s - f(x^k + s^k)| &\leq \mathcal{O}(\delta_k^3). \quad \text{w.p. } p_f \end{aligned}$$

$$p = p_f * p_g$$

Assumptions on models and estimates

For line search

$$\|\nabla f(x^k) - g_k\| \quad \leq \quad \mathcal{O}(\alpha_k \|g_k\|), \quad \text{w.p. } p_g$$

$$|f_k^0 - f(x^k)| \leq \mathcal{O}(\delta_k^2) \text{ and } |f_k^s - f(x^k + s^k)| \leq \mathcal{O}(\delta_k^2). \quad \text{w.p. } p_f$$

$$\mathbb{E}|f_k^0 - f(x^k)| \leq \mathcal{O}(\delta_k^2)$$

$$p = p_f * p_g$$

Stochastic TR: First-order convergence rate.

- Δ_k is the trust region radius.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\Delta_k^2$.
- $T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X^k)\| \leq \epsilon\}$.

Theorem

(Blanchet-Cartis-Menickelly-S. '17)

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{p}{2p-1}\left(\frac{L}{\epsilon^2}\right)\right),$$

Stochastic TR: Second-order convergence rate

- Δ_k is the trust region radius.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\Delta_k^2$.
- $T_\epsilon = \inf\{k \geq 0 : \max\{\|\nabla f(X^k)\|, -\lambda_{\min}(\nabla^2 f(X^k))\} \leq \epsilon\}$.

Theorem

(Blanchet-Cartis-Menickelly-S. '17)

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{p}{2p-1}\left(\frac{L}{\epsilon^3}\right)\right),$$

Stochastic line search: nonconvex case

- $\Delta_k = \alpha_k$ - the step size parameter.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^2 + (1-\nu)\theta\delta_k^2$.
- $T_\epsilon = \inf\{k \geq 0 : \|\nabla f(X^k)\| \leq \epsilon\}$.

Theorem

(Paquette-S. '18)

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{p}{2p-1}\left(\frac{L^3}{\epsilon^2}\right)\right),$$

Stochastic line search: convex case

- $\Delta_k = \alpha_k$ - the step size parameter.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^2 + (1-\nu)\theta\delta_k^2$.
- $T_\epsilon = \inf\{k \,:\, f(x_k) - f^* < \varepsilon\}$.
- $\Psi_k = \frac{1}{\nu\varepsilon} - \frac{1}{\Phi_k}$.

Theorem

(Paquette-S. '18)

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{p}{2p-1}\left(\frac{L^3}{\varepsilon}\right)\right),$$

Stochastic line search: strongly convex case

- $\Delta_k = \alpha_k$ - the step size parameter.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^2 + (1-\nu)\theta\delta_k^2$.
- $T_\epsilon = \inf\{k \,:\, f(x_k) - f^* < \varepsilon\}$.
- $\Psi_k = \log(\Phi_k) - \log(\nu\varepsilon)$.

Theorem

(Paquette-S. '18)

$$\mathbb{E}[T_\epsilon] \leq \mathcal{O}\left(\frac{p}{2p-1}\log\left(\frac{L^3}{\varepsilon}\right)\right),$$

Conclusions and Remarks

- We have a powerful framework based on bounding stoping time of a martingale which can be used to derive expected complexity bounds for adaptive stochastic methods.
- Accuracy requirement of function value estimates are much heavier than those for gradient estimates.
- Accuracy requirement of gradient value estimates are somewhat heavier than those for gradient estimates.
- Algorithms can converge even with constant probability of "iteration failure."

Depp Learning

The problem is not the problem. The problem is
your attitude about the problem

Jack Sparrow, Pirates of the Caribbean

Thanks for listening!

- J. Blanchet, C. Cartis, M. Menickelly, and K. Scheinberg, "Convergence Rate Analysis of a Stochastic Trust Region Method via Submartingales". *arXiv:1609.07428*, 2017.
- C. Paquette, K. Scheinberg, "A Stochastic Line Search Method with Convergence Rate Analysis". *arXiv:1807.07994*, 2018.