New Analysis of Adaptive Stochastic Optimization Methods via Supermartingales
Part II: Convergence analysis for stochastic line search

Courtney Paquette
Joint work with Katya Scheinberg

Waterloo

Lehigh University TRIPODS/DIMACS 2018
August 15, 2018

(Deterministic) Backtracking Line Search

Classical problem

$$\min_x \; f(x)$$

$f : \Omega \to \mathbf{R}$ is C^1 smooth w/ L-Lipschitz continuous gradient, bounded below

Gradient descent: $x_{k+1} = x_k - \alpha \nabla f(x_k), \quad \alpha \in (0, 1/L]$

(Deterministic) Backtracking Line Search

Classical problem

$$\min_{x} \; f(x)$$

$f : \Omega \to \mathbf{R}$ is C^1 smooth w/ L-Lipschitz continuous gradient, bounded below

Gradient descent: $x_{k+1} = x_k - \alpha \nabla f(x_k), \quad \alpha \in (0, 1/L]$

Backtracking Line Search Algorithm

- Compute $f(x_k)$ and $\nabla f(x_k)$
- Check sufficient decrease (Armijo '66)

$$\boxed{f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \theta \alpha_k \left\| \nabla f(x_k) \right\|^2}$$

- Successful: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ and $\alpha_{k+1} = \alpha_k$
- Unsuccessful: $x_{k+1} = x_k$ and $\alpha_k \downarrow$

$$\boxed{\text{Stepsize } \alpha_k \approx \frac{1}{L}}$$

Convergence Rate

Sufficient Decrease: $f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \theta \alpha_k \|\nabla f(x_k)\|^2$

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

Stochastic Line Search Question

Stochastic problem

$$\min_{x\in\mathbf{R}^n} f(x) = \mathbf{E}_\xi[\tilde{f}(x;\xi)], \qquad \xi \text{ is a random variable}$$

Examples

- *Empirical risk minimization*: ξ_i is a uniform r.v. over training set
- *More generally*: ξ is any sample or set of samples from data distribution

Stochastic Line Search Question

Stochastic problem

$$\min_{x \in \mathbf{R}^n} f(x) = \mathbf{E}_\xi[\tilde{f}(x;\xi)], \qquad \xi \text{ is a random variable}$$

Examples

- *Empirical risk minimization*: ξ_i is a uniform r.v. over training set
- *More generally*: ξ is any sample or set of samples from data distribution

(Stochastic) Backtracking Line Search Algorithm

- Compute stochastic estimates $\underbrace{g_k}_{\nabla f(x_k)}$, $\underbrace{f_k^0}_{f(x_k)}$, and $\underbrace{f_k^s}_{f(x_k - \alpha_k g_k)}$
- Check sufficient decrease (Armijo '66)

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

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

(Friedlander-Schmidt '12; Mahsereci-Hennig '17, ...)

$$f_k^s \le f_k^0 - \theta\alpha_k \|g_k\|^2 \Rightarrow f(x_k - \alpha_k g_k) \le f(x_k) - \theta\alpha_k \|g_k\|^2$$

Simple example: Know exact function values

$$f(x_{k+1}) \le f(x_k)$$

$$f_k^s \leq f_k^0 - \theta\alpha_k \|g_k\|^2 \Rightarrow f(x_k - \alpha_k g_k) \leq f(x_k) - \theta\alpha_k \|g_k\|^2$$

Simple example: Know exact function values

$$f(x_{k+1}) \leq f(x_k)$$

Challenges

- Bad function estimates may $\uparrow$ objective value

  Increase- $\alpha_k^2 \|g_k\|^2$

$$f_k^s \leq f_k^0 - \theta\alpha_k \|g_k\|^2 \Rightarrow f(x_k - \alpha_k g_k) \leq f(x_k) - \theta\alpha_k \|g_k\|^2$$

Simple example: Know exact function values

$$f(x_{k+1}) \leq f(x_k)$$

Challenges

- Bad function estimates may $\uparrow$ objective value

  Increase- $\alpha_k^2 \|g_k\|^2$

- Stepsizes, α_k, become arbitrarily small

$$f_k^s \leq f_k^0 - \theta\alpha_k \|g_k\|^2 \Rightarrow f(x_k - \alpha_k g_k) \leq f(x_k) - \theta\alpha_k \|g_k\|^2$$

Simple example: Know exact function values

$$f(x_{k+1}) \leq f(x_k)$$

Challenges

- Bad function estimates may ↑ objective value

  Increase- $\alpha_k^2 \|g_k\|^2$

- Stepsizes, α_k, become arbitrarily small

Question

Devise a line search for the stochastic problem with provable convergence guarantees using only knowable quantities.

Knowable quantities: e.g. bound on variance of $\nabla \tilde{f}$, $\tilde{f}$

(Bollapragada et al '17, Cartis-Scheinberg '17)

Proposed stochastic line search

Algorithm

- Compute random estimate of the gradient, g_k
- Compute random estimate of $f_k^0 \approx f(x_k)$ and $f_k^s \approx f(x_k - \alpha_k g_k)$
- Check the stochastic 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 $\alpha_k \|g_k\|^2 \geq \delta_k^2$, $\uparrow \delta_k$
  - Unreliable step: If $\alpha_k \|g_k\|^2 < \delta_k^2$, $\downarrow \delta_k$
- Unsucessful: $x_{k+1} = x_k$, $\alpha_k \downarrow$, and $\boxed{\delta_k \downarrow}$

What is δ_k?

Bad function estimates may $\uparrow$ objective value

$\alpha_k \|g_k\|$

$\delta \approx$ prediction of the size of $\alpha_k \|g_k\|$

$\approx$ size of a "trust region"

$\Rightarrow$ Largest $\uparrow$ in objective is at most δ_k^2

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

Stochastic Line Search

Algorithm

- Compute random estimate of the gradient, g_k
- Compute random estimate of $f_k^0 \approx f(x_k)$ and $f_k^s \approx f(x_k - \alpha_k g_k)$
- Check the stochastic 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 $\alpha_k \|g_k\|^2 \geq \delta_k^2$, $\uparrow\delta_k$
  - Unreliable step: If $\alpha_k \|g_k\|^2 < \delta_k^2$, $\downarrow\delta_k$
- Unsucessful: $x_{k+1} = x_k$, $\alpha_k \downarrow$, and $\boxed{\delta_k \downarrow}$

Randomness assumptions

- **Accurate gradient G_k w/ prob. p_g:**

$$\mathbf{Pr}(\|G_k - \nabla f(X_k)\| \leq \kappa_g \mathcal{A}_k \|G_k\| \mid \text{past}) \geq p_g$$

Randomness assumptions

- **Accurate gradient G_k w/ prob. p_g:**

$$\mathbf{Pr}(\|G_k - \nabla f(X_k)\| \leq \kappa_g \mathcal{A}_k \|G_k\| \mid \text{past}) \geq p_g$$

- **Accurate function estimates F_k^0 and F_k^s w/ prob. p_f:**

$$\begin{aligned}\mathbf{Pr}(&|f(X_k) - F_k^0| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \\ &\text{and} \quad |f(X_k - \mathcal{A}_k G_k) - F_k^s| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \mid \text{past}) \geq p_f\end{aligned}$$

Randomness assumptions

- **Accurate gradient G_k w/ prob. p_g:**

$$\mathbf{Pr}(\|G_k - \nabla f(X_k)\| \leq \kappa_g \mathcal{A}_k \|G_k\| \mid \text{past}) \geq p_g$$

- **Accurate function estimates F_k^0 and F_k^s w/ prob. p_f:**

$$\begin{aligned}\mathbf{Pr}(&|f(X_k) - F_k^0| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \\ \text{and} \quad & |f(X_k - \mathcal{A}_k G_k) - F_k^s| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \mid \text{past}) \geq p_f\end{aligned}$$

- **Variance condition**

$$\mathbf{E}[|F_k^0 - F(X_k)|^2 \mid \text{past}] \leq \theta^2 \Delta_k^4 \qquad \text{(same for } F_k^s\text{).}$$

Want to choose these probabilities (p_f, p_g) large enough

Randomness assumptions

- **Accurate gradient G_k w/ prob. p_g:**

$$\mathbf{Pr}(\|G_k - \nabla f(X_k)\| \leq \kappa_g \mathcal{A}_k \|G_k\| \mid \text{past}) \geq p_g$$

- **Accurate function estimates F_k^0 and F_k^s w/ prob. p_f:**

$$\mathbf{Pr}(|f(X_k) - F_k^0| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2$$
$$\text{and} \qquad |f(X_k - \mathcal{A}_k G_k) - F_k^s| \leq \varepsilon_f \mathcal{A}_k^2 \|G_k\|^2 \mid \text{past}) \geq p_f$$

- **Variance condition**

$$\mathbf{E}[|F_k^0 - F(X_k)|^2 \mid \text{past}] \leq \theta^2 \Delta_k^4 \qquad \text{(same for } F_k^s\text{)}.$$

Want to choose these probabilities (p_f, p_g) large enough

$p_f, p_g \geq 1/2$ at least, but p_f should be large.

Satisfying randomness assumptions

$$\min_{x \in \mathbf{R}^n} \; f(x) = \mathbf{E}_{\xi}[\tilde{f}(x;\xi)]$$

and bound on variance

$$\mathbf{E}(\|\nabla \tilde{f}(x,\xi_i) - \nabla f(x)\|^2) \leq V_g, \quad \mathbf{E}(|\tilde{f}(x;\xi_i) - f(x)|^2) \leq V_f.$$

Satisfying randomness assumptions

$$\min_{x \in \mathbf{R^n}} \; f(x) = \mathbf{E}_\xi[\tilde{f}(x;\xi)]$$

and bound on variance

$$\mathbf{E}(\|\nabla \tilde{f}(x,\xi_i) - \nabla f(x)\|^2) \le V_g, \quad \mathbf{E}(|\tilde{f}(x;\xi_i) - f(x)|^2) \le V_f.$$

Example: sampling

$$g_k = \frac{1}{|S_g|}\sum_{i \in S_g} \nabla f(x_k;\xi_i), \qquad f_k^0 = \frac{1}{|S_f|}\sum_{i \in S_f} f(x_k;\xi_i).$$

How many samples do we need?

Satisfying randomness assumptions

$$\min_{x\in\mathbf{R^n}} \; f(x) = \mathbf{E}_\xi[\tilde{f}(x;\xi)]$$

and bound on variance

$$\mathbf{E}(\|\nabla \tilde{f}(x,\xi_i) - \nabla f(x)\|^2) \le V_g, \quad \mathbf{E}(|\tilde{f}(x;\xi_i) - f(x)|^2) \le V_f.$$

Example: sampling

$$g_k = \frac{1}{|S_g|}\sum_{i\in S_g} \nabla f(x_k;\xi_i), \qquad f_k^0 = \frac{1}{|S_f|}\sum_{i\in S_f} f(x_k;\xi_i).$$

How many samples do we need?

Idea: Chebyshev Inequality

$$|S_g| \approx \tilde{O}\left(\frac{V_g}{\mathcal{A}_k^2\,\|G_k\|^2}\right), \qquad |S_f| \approx \tilde{O}\left(\max\left\{\frac{V_f}{\mathcal{A}_k^4\,\|G_k\|^4}, \frac{V_f}{\theta^2\Delta_k^4}\right\}\right)$$

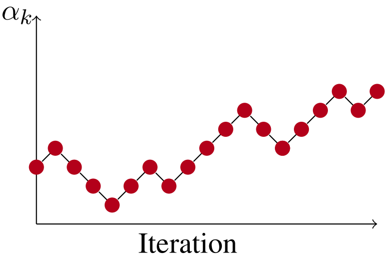
Dynamics of the stepsize

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success

Dynamics of the stepsize

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success
α_k bounded from 0	

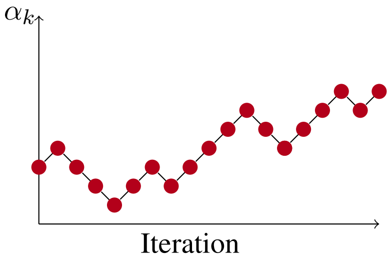
When $\alpha_k \lessapprox 1/L$, α_k move $\uparrow$ and $\downarrow$ like **random walk** with probability $p_g p_f$



Dynamics of the stepsize

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success
α_k bounded from 0	$\mathbf{Pr}(\limsup_k \mathcal{A}_k > 0) = 1$

When $\alpha_k \lesssim 1/L$, α_k move $\uparrow$ and $\downarrow$ like **random walk** with probability $p_g p_f$



Probability Viewpoint

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success

Probability Viewpoint

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success
α_k bounded from 0	$\mathbf{Pr}(\limsup_k \mathcal{A}_k > 0) = 1$

Probability Viewpoint

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success
α_k bounded from 0	$\mathbf{Pr}(\limsup_k \mathcal{A}_k > 0) = 1$
Function values decease each iteration	$\Phi_k \approx f(X_k) - f^*$ such that $\mathbf{E}[\Phi_{k+1} - \Phi_k \mid \text{past}] < 0$

Probability Viewpoint

Deterministic	Stochastic
$\alpha_k \leq 1/L \Rightarrow$ successful step	Good gradient/function estimates & stepsize $\leq 1/L$, $\Rightarrow$ success
α_k bounded from 0	$\mathbf{Pr}(\limsup_k \mathcal{A}_k > 0) = 1$
Function values decease each iteration	$\Phi_k \approx f(X_k) - f^*$ such that $\mathbf{E}[\Phi_{k+1} - \Phi_k \mid \text{past}] < 0$
Convergence rate: number of iterations until nearly optimal (e.g. $\|\nabla f(x)\| < \varepsilon$, $f(x) - f^* < \varepsilon$)	Convergence rate $\Rightarrow$ stopping times e.g. $T = \inf\{k > 0 \,:\, \|\nabla f(X_k)\| < \varepsilon\}$, $T = \inf\{k > 0 \,:\, f(X_k) - f^* < \varepsilon\}$

$\boxed{\text{Interested in } \mathbf{E}[T]}$

Renewal and reward process

Random process $\{\Phi_k, \mathcal{A}_k, W_k\}$

- $\Phi_k \in [0, \infty)$ and $\mathcal{A}_k \in [0, \infty)$
- W_k biased random walk with probability $p > 1/2$

$$\Pr(W_{k+1} = 1 | \text{ past}) = p \quad \text{and} \quad \Pr(W_{k+1} = -1 | \text{ past}) = 1 - p.$$

Renewal and reward process

Random process $\{\Phi_k, \mathcal{A}_k, W_k\}$

- $\Phi_k \in [0, \infty)$ and $\mathcal{A}_k \in [0, \infty)$
- W_k biased random walk with probability $p > 1/2$

$$\Pr(W_{k+1} = 1 | \text{past}) = p \quad \text{and} \quad \Pr(W_{k+1} = -1 | \text{past}) = 1 - p.$$

Assumptions

(i) $\exists \, \bar{\mathcal{A}}$ with

$$\mathcal{A}_{k+1} \geq \min \left\{ A_k e^{\lambda W_{k+1}}, \bar{\mathcal{A}} \right\}$$

(ii) $\exists$ nondecreasing $h : [0, \infty) \to (0, \infty)$ and constant Θ s.t.

$$\mathbf{E}[\Phi_{k+1} | \text{past}] \leq \Phi_k - \Theta h(\mathcal{A}_k).$$

Renewal and reward process

Random process $\{\Phi_k, \mathcal{A}_k, W_k\}$

- $\Phi_k \in [0, \infty)$ and $\mathcal{A}_k \in [0, \infty)$
- W_k biased random walk with probability $p > 1/2$

$$\Pr(W_{k+1} = 1 | \text{ past}) = p \quad \text{and} \quad \Pr(W_{k+1} = -1 | \text{ past}) = 1 - p.$$

Assumptions

(i) $\exists\, \bar{\mathcal{A}}$ with

$$\mathcal{A}_{k+1} \geq \min\left\{A_k e^{\lambda W_{k+1}}, \bar{\mathcal{A}}\right\}$$

(ii) $\exists$ nondecreasing $h : [0, \infty) \to (0, \infty)$ and constant Θ s.t.

$$\mathbf{E}[\Phi_{k+1} | \text{ past}] \leq \Phi_k - \Theta h(\mathcal{A}_k).$$

Thm: (Blanchet, Cartis, Menickelly, Scheinberg '17)

$$\mathbf{E}[T_\varepsilon] \leq \frac{p}{2p-1} \cdot \frac{\Phi_0}{\Theta h(\bar{\mathcal{A}})} + 1.$$

Convergence Result: Line search

Key observation

$$\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \left\|\nabla f(x_k)\right\|^2 + (1-\nu)\theta\delta_k^2$$

$$\begin{aligned}
\Rightarrow\ \Phi_{k+1} - \Phi_k = \nu(f(x_{k+1}) - f(x_k)) & \\
+ (1-\nu)\left(\alpha_{k+1}\left\|\nabla f(x_{k+1})\right\|^2 - \alpha_k\left\|\nabla f(x_k)\right\|^2\right) & \\
+ (1-\nu)\theta(\delta_{k+1}^2 - \delta_k^2) &
\end{aligned}$$

Convergence Result: Line search

Key observation

$$\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \left\|\nabla f(x_k)\right\|^2 + (1-\nu)\theta\delta_k^2$$

$$\begin{aligned}\Rightarrow\ \Phi_{k+1} - \Phi_k = \nu(f(x_{k+1}) - f(x_k))& \\ + (1-\nu)\left(\alpha_{k+1}\left\|\nabla f(x_{k+1})\right\|^2 - \alpha_k \left\|\nabla f(x_k)\right\|^2\right)& \\ + (1-\nu)\theta(\delta_{k+1}^2 - \delta_k^2)&\end{aligned}$$

Thm: (P-Scheinberg '18) If

$$p_g p_f > 1/2 \quad \text{and} \quad p_f \text{ sufficiently large,}$$

$$\mathbf{E}[\Phi_{k+1} - \Phi_k |\, \text{past}] \leq -\left(\mathcal{A}_k \left\|\nabla f(X_k)\right\|^2 + \theta\Delta_k^2\right)$$

Proof Idea:

- accurate gradient + accurate function estimates $\Rightarrow$ Φ_k *always* $\downarrow$
- all other cases $\Phi_k \uparrow$ by same amount

Convergence result, nonconvex

Stopping Time

$$T = \inf\{k \,:\, \|\nabla f(x_k)\| < \varepsilon\}$$

Convergence rate, nonconvex (P-Scheinberg '18)

If $p_g p_f > 1/2$ and p_f sufficiently large,

$$\mathbf{E}[T] \leq \mathcal{O}\left(\frac{1}{\varepsilon^2}\right).$$

Convex case

$$\min_{x \in \Omega} f(x) = \mathbf{E}[\tilde{f}(x, \xi)]$$

where

- f is convex and $\|\nabla f(x)\| \leq L_f$ for all $x \in \Omega$
- $\|x - x^*\| \leq D$ for all $x \in \Omega$

Stopping time: $T = \inf\{k : f(x_k) - f^* < \varepsilon\}$

Convex case

$$\min_{x\in\Omega} f(x) = \mathbf{E}[\tilde{f}(x,\xi)]$$

where

- f is convex and $\|\nabla f(x)\| \leq L_f$ for all $x \in \Omega$
- $\|x - x^*\| \leq D$ for all $x \in \Omega$

Stopping time: $T = \inf\{k \,:\, f(x_k) - f^* < \varepsilon\}$

Key observation:

$$\boxed{\Psi_k = \frac{1}{\nu\varepsilon} - \frac{1}{\Phi_k}}$$

(Convergence rate, convex) (P-Scheinberg '18)

If $p_g p_f > 1/2$ and p_f sufficiently large,

$$\mathbf{E}[T] \leq \mathcal{O}\left(\frac{1}{\varepsilon}\right)$$

Strongly convex case

$$\min_{x \in \Omega} f(x) = \mathbf{E}[\tilde{f}(x, \xi)]$$

where f is μ-strongly convex

Stopping Time: $T = \inf\{k \, : \, f(x_k) - f^* < \varepsilon\}$

Strongly convex case

$$\min_{x \in \Omega} f(x) = \mathbf{E}[\tilde{f}(x, \xi)]$$

where f is μ-strongly convex

Stopping Time: $T = \inf\{k \,:\, f(x_k) - f^* < \varepsilon\}$

Key observation:

$$\boxed{\Psi_k = \log(\Phi_k) - \log(\nu\varepsilon)}$$

Convergence rate, strongly convex (P-Scheinberg '18)

If $p_g p_f > 1/2$ and p_f sufficiently large,

$$\mathbf{E}[T] \leq \mathcal{O}\left(\log\left(\frac{1}{\varepsilon}\right)\right)$$

Thank You

References

Paquette, C. and Scheinberg, K. (2017).
A Stochastic Line Search Method with Convergence Rate Analysis.
arXiv: 1807.07994.