# An adaptive line search method for stochastic optimization

Courtney Paquette
Joint work with Katya Scheinberg

University of Waterloo

*Cornell ORIE Young Researcher Workshop 2018*
October 13, 2018

# (Deterministic) Backtracking Line Search

Classical problem

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

$f : \Omega \to \mathbf{R}$ with $L$-Lipschitz gradient

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

# (Deterministic) Backtracking Line Search

Classical problem

$$\min_{x \in \Omega} \ f(x)$$

$f : \Omega \to \mathbf{R}$ with $L$-Lipschitz gradient

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

**Backtracking Line Search Algorithm**

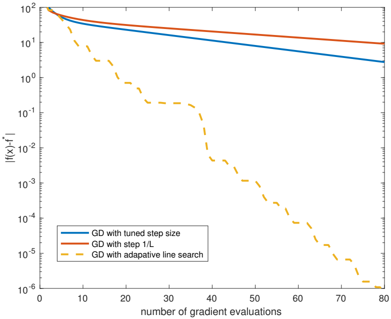
- Compute $f(x_k)$ and $\nabla f(x_k)$
- Check sufficient decrease (Armijo '66)

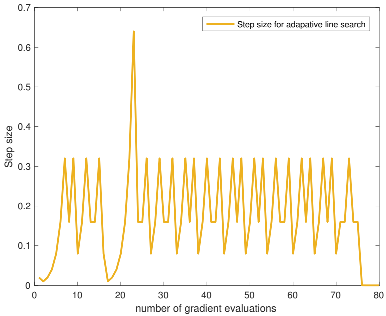
$$\boxed{f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \theta \alpha_k \left\|\nabla f(x_k)\right\|^2}$$

- Successful: $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ and increase $\alpha_k \Rightarrow \alpha_{k+1} = \gamma^{-1} \alpha_k$
- Unsuccessful: $x_{k+1} = x_k$ and decrease $\alpha_k \Rightarrow \alpha_{k+1} = \gamma \alpha_k$

# Motivation: Adaptivity (faster convergence)

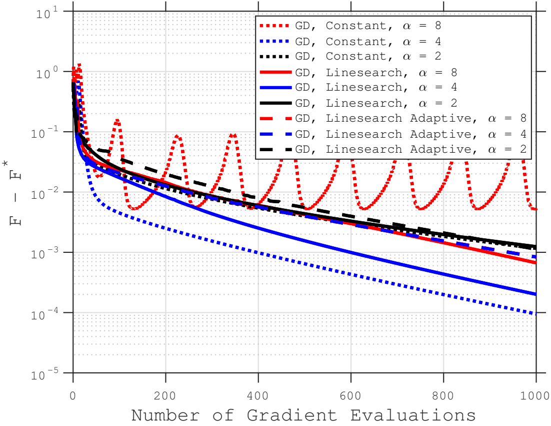
$$\min_x \; \frac{1}{2}x^T Ax - b^T x$$



# Motivation: Adaptivity (stability)

$$\min_{\theta} \; \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-y_i(\theta^T x_i)) + \frac{\lambda}{2}\|\theta\|_2^2$$



# Stochastic setting

Stochastic problem

$$\min_{x \in \Omega} f(x)$$

- $f : \Omega \to \mathbf{R}$ with $L$-Lipschitz gradients
- $f(x)$ is stochastic, given $x$ obtain estimate $\tilde{f}(x;\xi)$ and $\nabla \tilde{f}(x;\xi)$ where $\xi$ is random variable
- Central task in machine learning

$$f(x) = \mathbf{E}_{\xi \sim P}[\tilde{f}(x;\xi)]$$

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

**Question**

Can the line search technique be adapted to stochastic setting using only knowable quantities?

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

# Related works

## Subsampling and second-order methods

- Mahoney, Roosta-Khorasani, and Xu; "Newton-Type Methods for Non-convex optimization under inexact Hessian information" (2018)
- Tripuraneni, Stern, Jin, Reiger, and Jordan; "Stochastic cubic regularization for fast nonconvex optimization" (2017)
- Blanchet, Cartis, Menickelly, and Scheinberg; "Convergence rate analysis of a stochastic trust region method for nonconvex optimization" (2016)

## Line search & heuristics Previous work requires: $\nabla f(x)$, $\alpha_k \to 0$

- Bollapragada, Byrd, and Nocedal; "Adaptive sampling strategies for stochastic optimization" (to appear in SIOPT 2017)
- Friedlander and Schmidt; "Hybrid deterministic-stochastic methods for data fitting" (2012, SIAM Sci. Comput)
- Mahsereci and Hennig; "Probabilistic line search for stochastic optimization" (JMLR 2018; NIPS 2015)

# Stochastic backtracking line search

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

$$f_k^+ \leq f_k - \theta \alpha_k \|g_k\|^2$$

- Successful: $x_{k+1} = x_k - \alpha_k g_k$ and increase $\alpha_k \Rightarrow \alpha_{k+1} = \gamma^{-1} \alpha_k$
- Unsuccessful: $x_{k+1} = x_k$ and decrease $\alpha_k \Rightarrow \alpha_{k+1} = \gamma \alpha_k$

# Stochastic backtracking line search

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

$$f_k^+ \leq f_k - \theta \alpha_k \|g_k\|^2$$

- Successful: $x_{k+1} = x_k - \alpha_k g_k$ and increase $\alpha_k \Rightarrow \alpha_{k+1} = \gamma^{-1} \alpha_k$
- Unsuccessful: $x_{k+1} = x_k$ and decrease $\alpha_k \Rightarrow \alpha_{k+1} = \gamma \alpha_k$

**Challenges**

$$f_k^+ \leq f_k - \theta \alpha_k \|g_k\|^2 \quad \overset{??}{\Rightarrow} \quad f(x_k - \alpha_k g_k) \leq f(x_k) - \theta \alpha_k \|\nabla f(x_k)\|^2$$

- Bad function estimates may ↑ objective value

  Increase at most $\alpha_k^2 \|g_k\|^2$

- Stepsizes, $\alpha_k$, become arbitrarily small

# Stochastic line search

## Algorithm

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

$$f_k^+ \leq f_k - \theta \alpha_k \|g_k\|^2$$

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

  - Reliable step: If $\alpha_k \|g_k\|^2 \geq \delta_k^2$, $\uparrow \delta_k \Rightarrow \delta_{k+1}^2 = \gamma^{-1}\delta_k^2$
  - Unreliable step: If $\alpha_k \|g_k\|^2 < \delta_k^2$, $\downarrow \delta_k \Rightarrow \delta_{k+1}^2 = \gamma\delta_k^2$

- Unsucessful: $x_{k+1} = x_k$, decrease $\alpha_k$, and decrease $\delta_k$
  $\Rightarrow \alpha_{k+1} = \gamma\alpha_k$ and $\delta_{k+1}^2 = \gamma\delta_k^2$.

# Randomness assumptions

- **Accurate gradient $g_k$ w/ prob. $p_g$:**

$$\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \le \alpha_k \|g_k\| \mid \text{past}) \ge p_g$$

- **Accurate function estimates $f_k$ and $f_k^+$ w/ prob. $p_f$:**

$$\begin{aligned} \mathbf{Pr}(&|f(x_k) - f_k| \le \alpha_k^2 \|g_k\|^2 \\ &\text{and} \quad |f(x_k - \alpha_k g_k) - f_k^+| \le \alpha_k^2 \|g_k\|^2 \mid \text{past}) \ge p_f \end{aligned}$$

# Randomness assumptions

- **Accurate gradient $g_k$ w/ prob. $p_g$:**

$$\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \leq \alpha_k \|g_k\| \mid \text{past}) \geq p_g$$

- **Accurate function estimates $f_k$ and $f_k^+$ w/ prob. $p_f$:**

$$\begin{aligned}\mathbf{Pr}(&|f(x_k) - f_k| \leq \alpha_k^2 \|g_k\|^2 \\ \text{and} \quad &|f(x_k - \alpha_k g_k) - f_k^+| \leq \alpha_k^2 \|g_k\|^2 \mid \text{past}) \geq p_f\end{aligned}$$

- **Variance condition**

$$\mathbf{E}[|f_k - f(x_k)|^2 \mid \text{past}] \leq \theta^2 \delta_k^4 \qquad (\text{same for } f_k^+).$$

Question: How 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

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

and bound on variance

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

# Satisfying randomness assumptions

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

and bound on variance

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

**Example: sampling**

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

How many samples do we need?

# Satisfying randomness assumptions

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

and bound on variance

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

**Example: sampling**

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

How many samples do we need?

**Idea:** Chebyshev Inequality

$$|S_g| \approx \tilde{O}\left(\frac{V_g}{\alpha_k^2 \|g_k\|^2}\right), \qquad |S_f| \approx \tilde{O}\left(\max\left\{\frac{V_f}{\alpha_k^4 \|g_k\|^4}, \frac{V_f}{\delta_k^4}\right\}\right)$$

# Stochastic Process

- Random process $\{\Phi_k, \mathcal{A}_k\} \geq 0$
- Stopping time $T_\varepsilon$
- $W_k$ 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

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

# Stochastic Process

- Random process $\{\Phi_k, \mathcal{A}_k\} \geq 0$
- Stopping time $T_\varepsilon$
- $W_k$ 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

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

(ii) $\exists$ nondecreasing $h : [0, \infty) \to (0, \infty)$ such that

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

# Stochastic Process

- Random process $\{\Phi_k, \mathcal{A}_k\} \geq 0$
- Stopping time $T_\varepsilon$
- $W_k$ 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

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

(ii) $\exists$ nondecreasing $h : [0, \infty) \to (0, \infty)$ such that

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

**Optimization viewpoint**

- $\Phi_k$ is progress toward optimality
- $\mathcal{A}_k$ is step size parameter
- $T_\varepsilon$ is the first iteration $k$ to reach accuracy $\varepsilon$
- $\bar{\mathcal{A}} = 1/L$

# Stochastic process

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

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

Convergence result

$\mathbf{E}[T_\varepsilon]$ = expected number of iterations until reach accuracy $\varepsilon$

**Main idea of proof:**

- $\Phi_k$ is a supermartingale and $T_\varepsilon$ is a stopping time
- Compute expected number of times (renewals, $N(T_\varepsilon)$) $\mathcal{A}_k$ returns to $\bar{\mathcal{A}}$ before $T_\varepsilon$ (Wald's Identity)
- Optional stopping time relates expected renewals to supermartingale

# Convergence result: relationship to line search

## Key observations

- $\Phi_k = \underbrace{\nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^2}_{\text{balance each other}} + (1-\nu)\theta{\delta_k}^2$
- $\mathcal{A}_k = \alpha_k$, random walk with $p = p_g p_f$
- $T_\varepsilon = \inf\{k \geq 0 \,:\, \|\nabla f(x_k)\| < \varepsilon\}$
- $\bar{\mathcal{A}} = 1/L$

# Convergence result: relationship to line search

**Key observations**

- $\Phi_k = \underbrace{\nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^2}_{\text{balance each other}} + (1-\nu)\theta{\delta_k}^2$
- $\mathcal{A}_k = \alpha_k$, random walk with $p = p_g p_f$
- $T_\varepsilon = \inf\{k \geq 0 \,:\, \|\nabla f(x_k)\| < \varepsilon\}$
- $\bar{\mathcal{A}} = 1/L$

**Thm:** (P-Scheinberg '18) If

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

$$\mathbf{E}[\Phi_{k+1} - \Phi_k |\, \text{past}] \leq -\left(\alpha_k \|\nabla f(x_k)\|^2 + \theta\delta_k^2\right)$$

*Proof Idea:*

(1) accurate gradient + accurate function est. $\Rightarrow \Phi_k \downarrow$ by $\alpha_k \|\nabla f(x_k)\|^2$

(2) all other cases $\Phi_k \uparrow$ by $\alpha_k \|\nabla f(x_k)\|^2 + \theta\delta_k^2$

(3) Choose probabilities $p_f$, $p_g$ so that the (1) occurs more often

# Convergence result, nonconvex

**Stopping Time**

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

**Convergence rate, nonconvex** (P-Scheinberg '18)

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

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

# Convex case

**Assumptions:**

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

# Convex case

**Assumptions:**

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

**Key observation:**

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

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

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

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

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

# Strongly convex case

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

# Strongly convex case

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

**Key observation:**

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

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

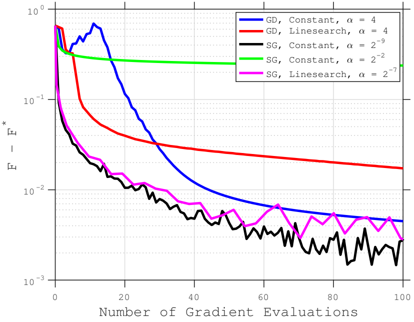
**Convergence rate, strongly convex** (P-Scheinberg '18)

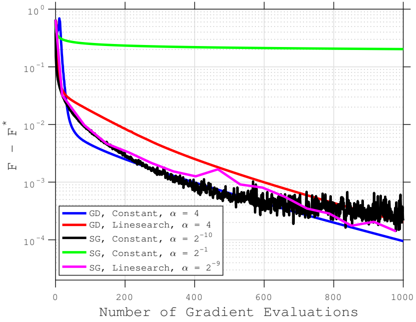
If $p_g p_f > 1/2$ and $p_f$ sufficiently large,

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

# Preliminary results

$$\min_{\theta} \; \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-y_i(\theta^T x_i))) + \frac{\lambda}{2} \|\theta\|_2^2$$



# Open questions and extensions

## Conclusions

- General framework for convergence results
- Convergence analysis (nonconvex, convex, and strongly convex) for a line search algorithm with gradient descent.

# Open questions and extensions

## Conclusions

- General framework for convergence results
- Convergence analysis (nonconvex, convex, and strongly convex) for a line search algorithm with gradient descent.

## Applications of the stochastic process

- Line search, trust region methods (Blanchet, Cartis, Menickelly, Scheinberg '17), and cubic regularization?
- Extensions into 2nd order stochastic methods with Hessian guarantees?

## Open problems

- Finding a good practical stochastic line search for machine learning; sampling procedure too conservative
- Extending line search procedure to stochastic BFGS

Thank You