Algorithms for stochastic problems lacking convexity or smoothness

Courtney Paquette

University of Waterloo

Google Tech Talk
January 16, 2019

Research Directions

(1). Thesis Work

- acceleration
- nonsmooth analysis of eigenvalues
- composite nonlinear models $(h \circ c)$
- statistical guarantees for nonconvex problems

(2). Post doc

- stochastic optimization
- constrained conjugate gradient

Research Directions

(1). Thesis Work

- acceleration
- nonsmooth analysis of eigenvalues
- composite nonlinear models $(h \circ c)$
- **statistical guarantees for nonconvex problems**

(2). Post doc

- **stochastic optimization**
- constrained conjugate gradient

(1). Local search for non-smooth and non-convex problems
(2). Adaptive line search for stochastic optimization

$\boxed{\text{Local search for non-smooth and non-convex problems}}$

Joint work with D. Davis, D. Drusvyatskiy, and K. MacPhee

Why study nonsmooth and nonconvex optimization?

$$\min_x \; g(x)$$

Nonsmooth and nonconvex losses arise often...

- Structure (sparsity), robustness (outliers), stability (better conditioning)

Common problem class: (convex) $\circ$ (smooth)

(Fletcher '80, Powell '83, Burke '85, Wright '90, Lewis-Wright '08, Cartis-Gould-Toint '11)

Global convergence guarantees for composite class

Drusvyatskiy-P '18; (Math. Program)

Local search

$$\min_{x} \; g(x), \qquad \left(e.g. \;\; g(x) = \sum_{i=1}^{m} g_i(x) \right)$$

Strategy:

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with a rapidly converging algorithm

Local search

$$\min_x \; g(x), \quad g \text{ is nonconvex and nonsmooth} \quad \left(e.g.\;\; g(x) = \sum_{i=1}^{m} g_i(x) \right)$$

Strategy:

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with a rapidly converging algorithm

Is there a generic gradient-based local search procedure for nonsmooth and nonconvex problems?

Local search

$$\min_x \; g(x)$$

Strategy:

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with $\boxed{\text{rapidly}}$ converging algorithm

Local search

$$\min_x \; g(x)$$

Strategy:

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with $\boxed{\text{rapidly}}$ converging algorithm

Gradient-based methods

convex + **regularity** $\Rightarrow$ rapid convergence

Local search

$$\min_x \; g(x)$$

Strategy:

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with $\boxed{\text{rapidly}}$ converging algorithm

Gradient-based methods

convex + **regularity** $\Rightarrow$ rapid convergence

Regularity condition

Sharpness: A function $g : \mathbb{R}^d \to \mathbb{R}$ is μ-sharp if

$$\boxed{g(x) - \min\; g \geq \mu \cdot \operatorname{dist}(x; S), \quad \text{for all } x \in \mathbb{R}^d}$$

where S is the set of minimizers of g.

Convergence rates:

- (Prox) gradient: sharpness + convexity $\Rightarrow$ quadratic
- Subgradient (Shor '77, 'Polyak 67): sharpness + convexity $\Rightarrow$ linear

Example: Robust Phase Retrieval

Problem: Find $x \in \mathbb{R}^d$ such that

$$(a_i^T x)^2 \approx b_i \quad a_1, \ldots, a_m \in \mathbb{R}^d, \quad b_1, \ldots, b_m \in \mathbb{R}.$$

Composite formulation:

$$\min_x \; g(x) := \frac{1}{m} \sum_{i=1}^m |(a_i^T x)^2 - b_i|$$

Assumptions: $a_i \sim N(0, I_d)$ independently and $b = (A\bar{x})^2$ for some $\bar{x} \in \mathbb{R}^d$.

Example: Robust Phase Retrieval

Problem: Find $x \in \mathbb{R}^d$ such that

$$(a_i^T x)^2 \approx b_i \quad a_1, \ldots, a_m \in \mathbb{R}^d, \quad b_1, \ldots, b_m \in \mathbb{R}.$$

Composite formulation:

$$\min_x \; g(x) := \frac{1}{m} \sum_{i=1}^{m} |(a_i^T x)^2 - b_i|$$

Assumptions: $a_i \sim N(0, I_d)$ independently and $b = (A\bar{x})^2$ for some $\bar{x} \in \mathbb{R}^d$.

Consequences: $\exists$ constants $\beta, \alpha > 0$ such that with probability $1 - e^{-cm}$

- **Weakly-convex**: (Duchi-Ruan '17)

$$y \mapsto g(y) + \frac{\rho}{2} \|y\|_2^2 \quad \text{is convex}$$

- **Sharpness**: (Eldar-Mendelson '14)

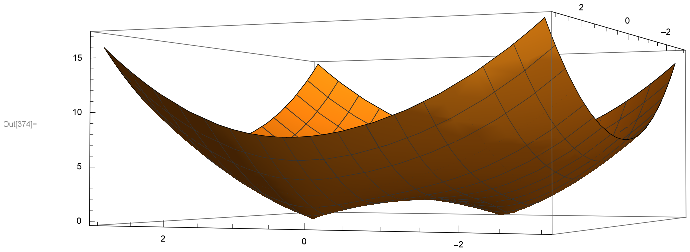
$$g(x) \geq \alpha \|\bar{x}\|_2 \operatorname{dist}(x, \{\pm \bar{x}\}).$$

Holds even when 1/2 the points are corrupted!

Intuition

g approximates the population objective:

$$g_P(x) = \mathbf{E}_{a\sim N}[|\langle a, x\rangle^2 - \langle a, \bar{x}\rangle^2|]$$



Good neighborhood

$$\min_{x} \; g(x), \qquad \text{where } g \text{ is } \boldsymbol{\mu}\text{-sharp and } \boldsymbol{\rho}\text{-weakly convex.}$$

- (convex) $\circ$ (smooth) structure always weakly-convex

Local Search Procedure

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with a rapidly converging algorithm

Good neighborhood

$$\min_x \; g(x), \qquad \text{where } g \text{ is } \boldsymbol{\mu}\text{-sharp and } \boldsymbol{\rho}\text{-weakly convex.}$$

- (convex) $\circ$ (smooth) structure always weakly-convex

Local Search Procedure

- Find a moderately accurate solution $\hat{x}$ at a low sample complexity cost
- Refine $\hat{x}$ with a rapidly converging algorithm

Lemma (Davis-Drusvyatskiy-MacPhee-**P**)

No extraneous stationary points of g lie in the tube:

$$\boxed{\mathcal{T} := \left\{ x \in \mathbb{R}^d \, : \, \operatorname{dist}(x; S) < \frac{\boldsymbol{\mu}}{\boldsymbol{\rho}} \right\}}$$

"Lipschitz" constant: $\qquad L := \sup \left\{ \|\xi\| \, : \, \xi \in \partial g(x), x \in \mathcal{T} \right\}.$

$\boldsymbol{\kappa} = \frac{\boldsymbol{L}}{\boldsymbol{\mu}}$ acts like the "condition" number

Eg.: phase retrieval

- spectral initialization (Wang et al. '16, Duchi-Ruan '17)

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly

Meta-Theorem:

Simple algorithms for sharp and weakly convex functions converge rapidly

Polyak subgradient method:

$$x^+ = x - \left(\frac{g(x) - \inf\ g}{\|v\|^2}\right) v \quad \text{where } v \in \partial g(x).$$

Thm: (Polyak '67, Davis-Drusvyatskiy-MacPhee-**P** '17)

Suppose that g is

- ρ-weakly convex (meaning $g + \frac{\rho}{2}\|\cdot\|^2$ is convex)
- L-Lipschitz
- μ-sharp
- $\mathrm{dist}(x_0, S) \leq \frac{\mu}{2\rho}$

Then

$$\frac{\mathrm{dist}(x_{k+1}, S)}{\mathrm{dist}(x_k, S)} \leq \sqrt{1 - \left(\frac{\mu}{L\sqrt{2}}\right)^2}, \qquad \text{for all } k.$$

Eg: phase retrieval

- $\frac{\mu}{\rho}$, $\frac{\mu}{L}$ are **dimension independent** w.h.p. (Eldar-Mendelson '14)

Subgradient methods

What happens when $\inf\ g$ is unknown?

Subgradient method geometrically decaying stepsize:

$$x_{t+1} = x_t - \left(\sqrt{\left(1 - \left(\tfrac{\mu}{L}\right)^2\right)}\right)^t \frac{v_t}{\|v_t\|} \quad \text{where } v \in \partial g(x).$$

Thm: (Goffin '77, Shor, Davis-Drusvyatskiy-MacPhee-P '17)

Suppose g is

- ρ-weakly convex
- L-Lipschitz, μ-sharp
- $\mathrm{dist}(x_0, S) < \frac{\mu}{\rho}$

Then,

$$\mathrm{dist}^2(x_t, S) \leq \frac{\mu^2}{\rho^2}\left(1 - \left(\frac{\mu}{L}\right)^2\right)^t$$

Numerical Experiments

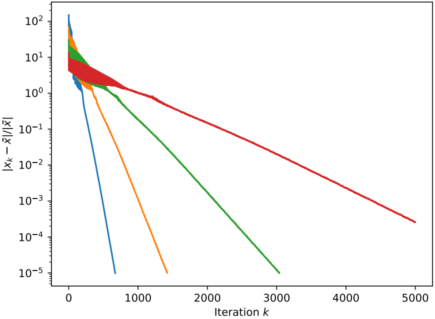


Figure: Subgradient geometric decaying: Robust phase retrieval

Other examples

- **Robust PCA** (Candes et al. ’11, Chandrasekaran et al. ’11, Netrapalli et al. ’14)

$$\min_{X\in\mathbb{R}^{d\times r}, Y\in\mathbb{R}^{r\times k}} \quad \|XY - D\|_1$$

- **Blind deconvolution/bi-convex sensing** (Ling-Strohmer ’15, Ahmed et al. ’14)

$$\min_{x,w} \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, w\rangle\langle r_i, x\rangle - b_i|$$

- **Covariance Estimation** (Chen et. al ’15, Davis-Drusvyatskiy-MacPhee-**P** ’18)

$$\min_{x} \frac{1}{m}\sum_{i=1}^{m} |\langle XX^T, a_{2i}a_{2i}^T - a_{2i-1}a_{2i-1}^T\rangle - (b_{2i} - b_{2i-1})|$$

- **conditional value-at-risk, dictionary learning, group synchronization,...**

Open questions and extensions

Conclusions

- local search procedure for nonsmooth, nonconvex problems
- Statistical well-posedness $\Rightarrow$ good initialization strategies and regularity

Examples

- Robust phase retrieval, covariance estimation, blind deconvolution...
- Matrix factorization?? Robust PCA??

Extensions

- Stochastic variants with rates in expectation (Davis-Drusvyatskiy-**P** '17, Duchi-Ruan '17, Davis-Drusvyatskiy '18)
- Bregman divergences (measure sharpness/Lipschitz w.r.t. norm other than $\|\cdot\|^2$) (Davis-Drusvyatskiy-MacPhee '18)

Adaptive line search method
for smooth stochastic optimization

Joint work with K. Scheinberg

Stochastic optimization

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

Stochastic gradient descent (SGD):

$$x_{k+1} \leftarrow x_k - \alpha g_k \qquad \text{where } g_k = \nabla \tilde{f}(x_k;\xi)$$

- **Major drawback:** stepsize, α, requires lots of tuning

Stochastic optimization

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

Stochastic gradient descent (SGD):

$$x_{k+1} \leftarrow x_k - \alpha g_k \qquad \text{where } g_k = \nabla \tilde{f}(x_k;\xi)$$

- **Major drawback:** stepsize, α, requires lots of tuning

Deterministic setting: Use line search techniques

Question:
Can the line search technique be adapted to the stochastic setting?

(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 $\alpha_k \Rightarrow \alpha_{k+1} = \gamma \alpha_k$

(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)

$$\boxed{\underbrace{f(x_k - \alpha_k \nabla f(x_k))}_{\text{function value at next step}} \leq \underbrace{f(x_k) - \theta \alpha_k \left\| \nabla f(x_k) \right\|^2}_{\text{linearization of } f \text{ at current step}}}$$

- 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 $\alpha_k \Rightarrow \alpha_{k+1} = \gamma \alpha_k$

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 ξ is random variable
- Central task in machine learning

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

  - *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

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

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, α_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 α_k, and $\boxed{\text{decrease } \delta_k}$
  $\Rightarrow \alpha_{k+1} = \gamma \alpha_k$ and $\delta_{k+1}^2 = \gamma \delta_k^2$.

Randomness assumptions

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

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

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

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

Randomness assumptions

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

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

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

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

- **Variance condition**

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

$$\boxed{\text{Question: How to choose these probabilities } (p_f, p_g) \text{ large enough?}}$$

$p_f, p_g \geq 1/2$ 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

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

Example: sampling

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

$$\boxed{\text{How many samples do we need?}}$$

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?

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_ε
- 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_ε
- 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_ε
- 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

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

Stochastic process

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

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

Convergence result

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

Main idea of proof:

- Φ_k is a supermartingale and T_ε is a stopping time
- Compute expected number of times (renewals, $N(T_\varepsilon)$) $\mathcal{A}_k$ returns to $\bar{\mathcal{A}}$ before T_ε (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,

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

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

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

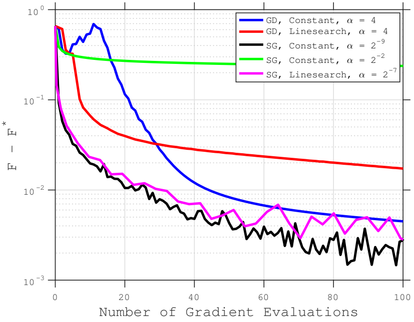
Convergence rate, strongly convex (P-Scheinberg '18)

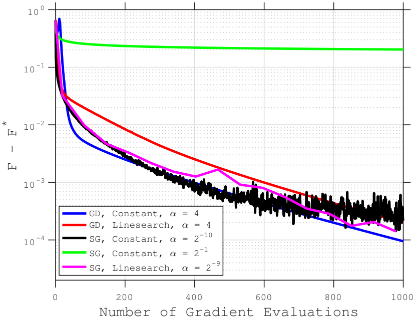
If $p_g p_f > 1/2$ and p_f sufficiently large,

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

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 Wolfe conditions (BFGS)

References

Davis, D., Drusvyatskiy, D., MacPhee, K., and Paquette, C. (2018).
Subgradient methods for sharp, weakly convex functions.
J. Optim. Theory App.

Davis, D., Drusvyatskiy, D., and Paquette, C. (2017).
The nonsmooth landscape of phase retrieval.
arXiv:1711.03247.

Drusvyatskiy, D. and Paquette, C. (2018).
Efficiency of minimizing compositions of convex functions and smooth maps.
Math. Program.

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