Complexity analysis framework of adaptive stochastic optimization methods via martingales.

Katya Scheinberg

joint work with J. Blanchet (Stanford), C. Cartis (Oxford), F. Curtis (Lehigh), M. Menickelly (Argonne) and C. Paquette (McGill)

July, 15 2020

Cornell University
Operations Research and
Information Engineering

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
- Parameter dependency - most methods require tuning of step size and other parameters
- Complexity - emphasis on complexity bounds

Unconstrained Optimization

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

- We will assume throughout that f is sufficiently smooth.
- f is nonconvex unless specified.
- When $f(x)$ is deterministic, standard methods are 1. line search, 2. trust region and 3. cubicly regularized Newton.
- When $f(x)$ is stochastic, standard method is stochastic gradient descent and variants.
- When $f(x)$ has biased noise and/or no derivative information, we use other methods (e.g. black box optimization).
- Can line search, trust region and regularized Newton methods be used in nondeterministic settings?

Generic Adaptive Deterministic Method

0. Initialization

   Choose constants $\eta \in (0,1)$, $\gamma \in (1,\infty)$, and $\overline{\alpha} \in (0,\infty)$. Choose an initial iterate $x_0 \in \mathbb{R}^n$ and stepsize parameter $\alpha_0 \in (0, \overline{\alpha}]$.

1. Determine model and compute step

   Choose a local model m_k of f around x_k. Compute a step $s_k(\alpha_k)$ such that the model reduction $m_k(x_k) - m_k(x_k + s_k(\alpha_k)) \geq 0$ is sufficiently large.

2. Check for sufficient reduction in f

   Check if $f(x_k) - f(x_k + s_k(\alpha_k))$ is sufficiently large relative to $m_k(x_k) - m_k(x_k + s_k(\alpha_k))$ using a condition parameterized by η.

3. Successful iteration

   If true (along with other potential requirements), then set $x_{k+1} \leftarrow x_k + s_k(\alpha_k)$ and $\alpha_{k+1} \leftarrow \min\{\gamma\alpha_k, \overline{\alpha}\}$.

4. Unsuccessful iteration

   Otherwise, $x_{k+1} \leftarrow x_k$ and $\alpha_{k+1} \leftarrow \gamma^{-1}\alpha_k$.

5. Next iteration

   Set $k \leftarrow k + 1$.

Particular Methods

For line search method

- $m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H s,\ H \succ 0$
- $s_k(\alpha_k) = -\alpha_k H^{-1} \nabla f(x_k)$
- Sufficient reduction: $f(x_k) - f(x_k + s_k(\alpha_k)) \geq -\eta \nabla f(x_k)^T s_k(\alpha_k)$

For trust region method

- $m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H s,\ H \sim \nabla^2 f(x_k)$
- $s_k(\alpha_k) = \arg\min_{s:\ \|s\| \leq \alpha_k} m_k(x_k + s)$
- Sufficient reduction: $\frac{f(x_k) - f(x_k + s_k(\alpha_k))}{m_k(x_k) - m_k(x_k + s_k(\alpha_k))} \geq \eta$

For cubicly regularized Newton method

- $m_k(x_k + s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s + \frac{1}{3\alpha_k} \|s\|^3,$
- $s_k(\alpha_k) = \arg\min_s m_k(x_k + s)$
- Sufficient reduction: $\frac{f(x_k) - f(x_k + s_k(\alpha_k))}{m_k(x_k) - m_k(x_k + s_k(\alpha_k))} \geq \eta$

What can happen?

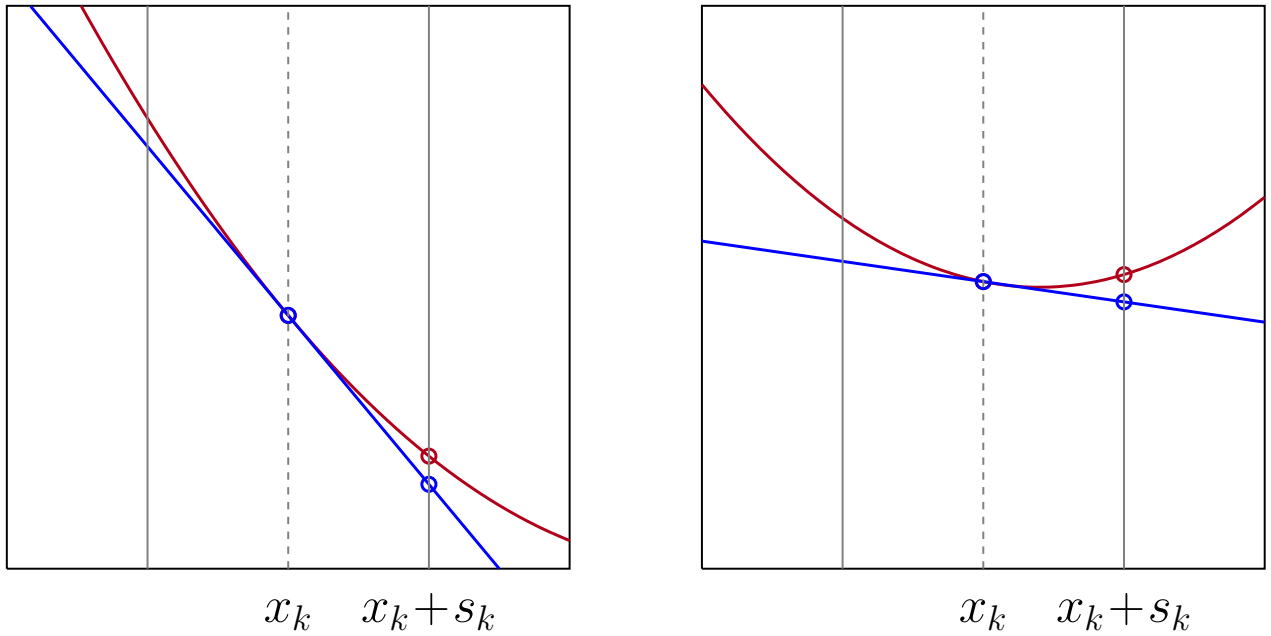


Figure: Illustration of successful (left) and unsuccessful (right) steps in a trust region method.

Framework for Convergence Rate Analysis

- $\{\Phi_k\} \geq 0$ - a sequence whose role is to measure progress of the algorithm.
- $\{W_k\}$ is a sequence of indicators; specifically, for all $k \in \mathbb{N}$, if iteration k is successful, then $W_k = 1$, and $W_k = -1$ otherwise.
- $\{\alpha_k\} \geq 0$ - a sequence of step size parameter values that obey the rule $\alpha_{k+1} = \min\{\gamma^{W_k}\alpha_k, \bar{\alpha}\}$
- T_ε, the *stopping time*, is the index of the first iterate that satisfies a desired convergence criterion parameterized by ε.

Condition 1

The following statements hold with respect to $\{(\Phi_k, \alpha_k, W_k)\}$ and T_ε.

1. There exists a scalar $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that for each $k \in \mathbb{N}$ such that $\alpha_k \leq \gamma\underline{\alpha}_\varepsilon$ implies $W_k = 1$. Therefore, $\alpha_k \geq \underline{\alpha}_\varepsilon$ for all $k \in \mathbb{N}$.
2. There exists a nondecreasing function $h_\varepsilon : [0, \infty) \to (0, \infty)$ and scalar $\Theta \in (0, \infty)$ such that, for all $k < T_\varepsilon$, $\Phi_k - \Phi_{k+1} \geq \Theta h_\varepsilon(\alpha_k)$.

Under Condition 1

$$T_\varepsilon \leq \frac{\Phi_0}{\Theta h_\varepsilon(\underline{\alpha}_\varepsilon)}.$$

Specifics of the analysis for the line search method

- $T_\varepsilon := \min\{k \in \mathbb{N} : \|\nabla f(x_k)\| \leq \varepsilon\}$,
- $\Phi_k := \nu(f(x_k) - f_*) + (1-\nu)\alpha_k\|\nabla f(x_k)\|^2$
  for some $\nu \in (0,1)$ (to be determined).
- If iteration k is successful ($W_k = 1$), then

$$f(x_k) - f(x_{k+1}) \geq \eta c_1 \alpha_k \|\nabla f(x_k)\|^2$$

- On unsuccessful iterations ($W_k = -1$), since no step is taken

$$\Phi_k - \Phi_{k+1} \geq (1-\nu)(1-\gamma^{-1})\alpha_k\|\nabla f(x_k)\|^2.$$

  Selecting ν sufficiently close to 1, while $\|\nabla f(x_k)\| > \varepsilon$, ensures Condition 1 with $h_\varepsilon(\alpha_k) := \alpha_k\varepsilon^2$ and $\Theta := (1-\nu)(1-\gamma^{-1})$.
- Thus,
  $T_\varepsilon \leq \frac{\nu(f(x_0)-f_*)+(1-\nu)\alpha_0\|\nabla f(x_0)\|^2}{(1-\nu)(1-\gamma^{-1})\underline{\alpha}\varepsilon^2} \Rightarrow T_\varepsilon = \mathcal{O}(\varepsilon^{-2})$.

Specifics of the analysis for the trust region method

- $T_\varepsilon := \min\{k \in \mathbb{N} : \|\nabla f(x_k)\| \leq \varepsilon\}$,
- $\Phi_k := \nu(f(x_k) - f_*) + (1-\nu)\alpha_k^2$
  for some $\nu \in (0,1)$ (to be determined).
- On successful iterations, $(W_k = 1)$, for some $c_2 \in (0, \infty)$,

$$f(x_k) - f(x_{k+1}) \geq \nu\eta c_2 \alpha_k^2$$

- On unsuccessful iterations $(W_k = -1)$, since no step is taken

$$\Phi_k - \Phi_{k+1} = (1-\nu)(1-\gamma^{-2})\alpha_k^2.$$

  Choosing ν sufficiently close to 1, while $\|\nabla f(x_k)\| > \varepsilon$, ensures Condition 1 with $h_\varepsilon(\alpha_k) := \alpha_k^2$ and $\Theta := (1-\nu)(1-\gamma^{-2})$.
- Thus, $T_\varepsilon \leq \frac{\nu(f(x_0)-f_*)+(1-\nu)\alpha_0^2}{(1-\nu)(1-\gamma^{-2})c_1^2\varepsilon^2} \Rightarrow T_\varepsilon = \mathcal{O}(\varepsilon^{-2})$.

Specifics of the analysis for the regularized Newton method

- $T_\varepsilon := \min\{k \in \mathbb{N} : \|\nabla f(x_{k+1})\| \leq \varepsilon\}$
- $\Phi_k := \nu(f(x_k) - f_*) + (1-\nu)\alpha_k \|\nabla f(x_k)\|^{3/2}$
  for some $\nu \in (0,1)$ (to be determined).
- If iteration k is successful ($W_k = 1$), for some $c_4 \in (0, \infty)$,

$$f(x_k) - f(x_{k+1}) \geq \eta c_4 \alpha_k \|\nabla f(x_{k+1})\|^{3/2}$$

- Otherwise ($W_k = -1$), since no step is taken

$$\Phi_k - \Phi_{k+1} \geq (1-\nu)(1-\gamma^{-1})\alpha_k \|\nabla f(x_{k+1})\|^{3/2}$$

- Choosing ν sufficiently close to 1 while $\min\{k \in \mathbb{N} : \|\nabla f(x_{k+1})\| > \varepsilon\}$, ensures required Condition 1 with $h_\varepsilon(\alpha_k) := \alpha_k \varepsilon^{3/2}$ and $\Theta := (1-\nu_\varepsilon)(1-\gamma^{-1})$.
- Thus $T_\varepsilon \leq \frac{\nu(f(x_0)-f_*)+(1-\nu)\alpha_0\|\nabla f(x_0)\|^{3/2}}{(1-\nu)(1-\gamma^{-1})\underline{\alpha}\varepsilon^{3/2}} \Rightarrow T_\varepsilon = \mathcal{O}(\varepsilon^{-3/2})$.

Other examples

Second order trust region method

- $T_\varepsilon := \min\{k \in \mathbb{N} : \chi_k \leq \varepsilon\}$,
  where $\chi_k := \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\}$
- $\Phi_k := \nu(f(x_k) - f_*) + (1-\nu)\alpha_k^3$
- $T_\varepsilon \leq \mathcal{O}(\varepsilon^{-3})$

Line search for convex functions

- $T_\varepsilon := \min\{k \in \mathbb{N} : f(x_k) - f_* \leq \varepsilon\}$
- $\Phi_k := \nu\left(\frac{1}{\varepsilon} - \frac{1}{f(x_k)-f_*}\right) + (1-\nu)\alpha_k$
- $T_\varepsilon \leq \mathcal{O}(\varepsilon^{-1})$

Similarly, line search for strongly convex functions

- $T_\varepsilon := \min\{k \in \mathbb{N} : f(x_k) - f_* \leq \varepsilon\}$
- $\Phi_k := \nu\left(\log\left(\frac{1}{\varepsilon}\right) - \log\left(\frac{1}{f(x_k)-f_*}\right)\right) + (1-\nu)\log(\alpha_k)$
- $T_\varepsilon \leq \mathcal{O}(\log(\varepsilon^{-1}))$

Generic Adaptive Stochastic Method

Initialization

Choose constants $\eta \in (0,1)$, $\gamma \in (1,\infty)$, and $\overline{\alpha} \in (0,\infty)$. Choose an initial iterate $x_0 \in \mathbb{R}n$ and stepsize parameter $\alpha_0 \in (0,\overline{\alpha}]$.

1. **Determine model and compute step**

   Choose a random local model m_k of f around x_k. Compute a step $s_k(\alpha_k)$ such that the model reduction $m_k(x_k) - m_k(x_k + s_k(\alpha_k)) \geq 0$ is sufficiently large.

2. **Check for sufficient reduction in f**

   Compute estimates $f_k^0 \sim f(x_k)$ and $f_k^s \sim f(x_k + s_k(\alpha_k))$ and check if $f_k^0 - f_k^s$ is sufficiently large relative to $m_k(x_k) - m_k(x_k + s_k(\alpha_k))$ using a condition parameterized by η.

3. **Successful iteration**

   If true (along with other potential requirements), then set $x_{k+1} \leftarrow x_k + s_k(\alpha_k)$ and $\alpha_{k+1} \leftarrow \min\{\gamma\alpha_k, \overline{\alpha}\}$.

4. **Unsuccessful iteration**

   Otherwise, $x_{k+1} \leftarrow x_k$ and $\alpha_{k+1} \leftarrow \gamma^{-1}\alpha_k$.

5. **Next iteration**

   Set $k \leftarrow k+1$.

Assumptions on the models/estimates

Model involves random/noisy estimates $f_k \sim f(x_k)$, $g_k \sim \nabla f(x_k)$ and $H_k \sim \nabla^2 f(x_k)$.
Different types of assumptions used in literature:

- $\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \leq \kappa\varepsilon \mid \text{ past}) \geq p_g$ - nonadaptive, strong
- $\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \leq \kappa\alpha_k \mid \text{ past}) \geq p_g$ - adaptive
- $\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \leq \kappa\alpha_k \|g_k\| \mid \text{ past}) \geq p_g$ - adaptive
- $\mathbf{Pr}(\|g_k - \nabla f(x_k)\| \leq \theta\|\nabla f(x_k)\| \mid \text{ past}) \geq p_g$ - not realizable.
- $\mathbf{Pr}(|f(x_k) - f_k^0| \leq \varepsilon^2 | \text{ past}) \geq p_f$ - nonadaptive, strong
  $\mathbf{Pr}(|f(x_k) - f_k^0| \leq \epsilon_f\alpha_k^2 | \text{ past}) \geq p_f$ - adaptive
- $\mathbf{Pr}(|f(x_k) - f_k^0| \leq \epsilon_f\alpha_k^2 \|g_k\|^2 | \text{ past}) \geq p_f$ -adaptive
- $\mathbf{Pr}(|f(x_k) - f_k^0| \leq \epsilon_f) = 1$ - nonadaptive, relaxed.

etc...

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

Why these assumptions?

- Gaussian smoothed gradients (Nesterov, Spokoiny 2017, Berahas Cao and S. 2020)

$$g(x) = \frac{1}{N\sigma}\sum_{i=1}^{N}(f(x+\sigma u_i) - f(x))u_i$$

with "high" probability

$$\|g(x) - \nabla f(x)\| \leq \theta\|\nabla f(x)\| + \mathcal{O}(\sigma)$$

- Stochastic gradient and Hessian

$$g(x) = \frac{1}{N}\sum_{i=1}^{N}(\nabla f_i(x))$$

bounded variance

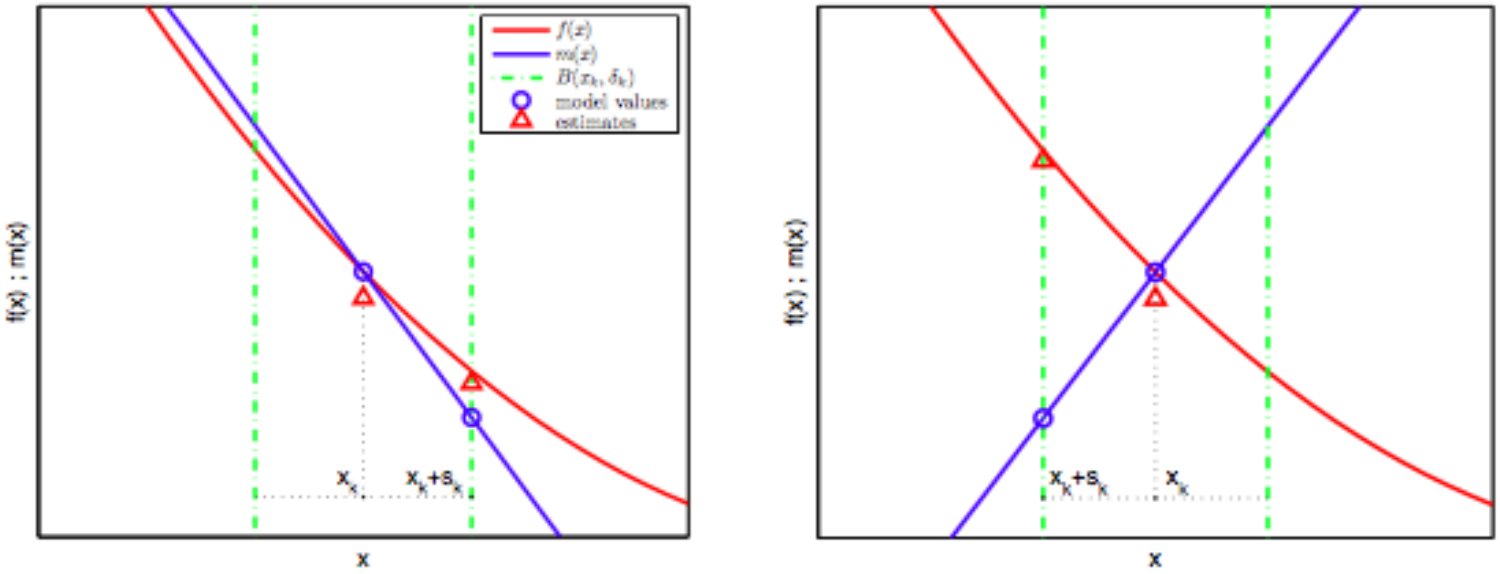
$$\mathbb{E}[\|g(x) - \nabla f(x)\|] \leq \theta\alpha$$

- Robust estimates in the presence of outliers (e.g. Yin, Chen, Ramchandran and Bartlett, 2018), with "high" probability

$$\|g(x) - \nabla f(x)\| \leq \theta\alpha$$

- Hessian sketching, sparse Hessian recovery, etc.

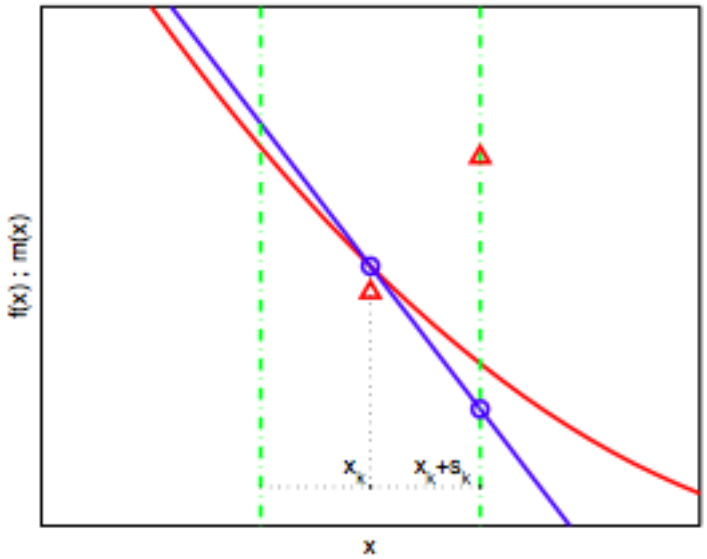
What can happen under random models?



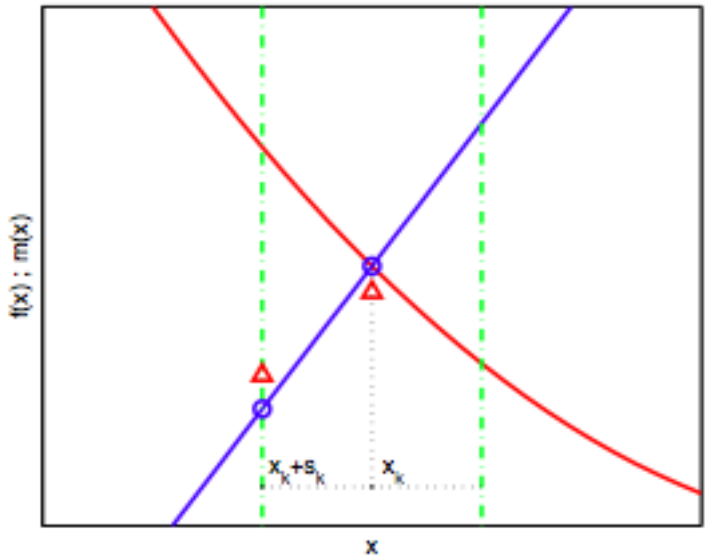
(a) Good model; good estimates.
True successful steps.

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

What else can happen under random estimates



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



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

Casting the Algorithm as a Stochastic Process

- $\{\Phi_k\} \geq 0$ - a random sequence whose role is to measure progress of the algorithm.
- $\{W_k\}$ is a sequence of random indicators; specifically, for all $k \in \mathbb{N}$, if iteration k is successful, then $W_k = 1$, and $W_k = -1$ otherwise.
- $\{\alpha_k\} \geq 0$ - a random sequence of step size parameter values that obey the rule $\alpha_{k+1} = \min\{\gamma^{W_k}\alpha_k, \bar{\alpha}\}$
- T_ε, the random *stopping time*, is the index of the first iterate that satisfies a desired convergence criterion parameterized by ε.

$\{\Phi_k, \alpha_k, W_k\}$ is a stochastic process and T_ϵ is its stopping time.

Recall Condition 1

The statements in red no longer hold with respect to $\{(\Phi_k, \alpha_k, W_k)\}$ and T_ε.

1. There exists a scalar $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that for each $k \in \mathbb{N}$ such that $\alpha_k \leq \gamma\underline{\alpha}_\varepsilon$, the iteration is guaranteed to be successful, i.e., $W_k = 1$. Therefore, $\alpha_k \geq \underline{\alpha}_\varepsilon$ for all $k \in \mathbb{N}$.
2. There exists a nondecreasing function $h_\varepsilon : [0, \infty) \to (0, \infty)$ and scalar $\Theta \in (0, \infty)$ such that, for all $k < T_\varepsilon$, $\Phi_k - \Phi_{k+1} \geq \Theta h_\varepsilon(\alpha_k)$.

The α_k Process

Condition 2 (i)

There exists a constant $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that, for $k < T_\varepsilon$

$$\alpha_{k+1} \geq \min(\gamma^{W_k} \alpha_k, \underline{\alpha}_\varepsilon),$$

where W_k is a random walk with positive drift (i.e. w.p $p > \frac{1}{2}$) $W_k = 1$)

Condition 2(ii)

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

$$\mathbb{E}(\Phi_{k+1} | \text{ past}) \leq \Phi_k - \Theta h(\alpha_k).$$

Bounding expected stopping time

Main Idea: This is a renewal-reward process and Φ_k is a supermartingale - $\mathbb{E}[\Phi_{k+1}|\,\text{past}] \leq \Phi_k - \Theta h_\varepsilon(\alpha_k)$ and, thus, $\Phi_0 \geq \Theta\mathbb{E}[\sum_{i=0}^{T_\epsilon} h(\alpha_i)]$.

- 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 Condition 2 hold. Then

$$\mathbb{E}[T_\varepsilon] \leq \frac{p}{2p-1} \cdot \frac{\Phi_0}{\Theta h(\underline{\alpha}_\varepsilon)} + 1.$$

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

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

$$p = p_f * p_g$$

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.

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

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

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

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

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

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

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

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

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

Theorem

(Paquette-S. '18)

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

Cubicly regularized Newton

- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k \left\|\nabla f(x_k)\right\|^{3/2} + ???.$
- $T_\epsilon = \inf\{k \,:\, \|\nabla f(x_{k+1})\| < \varepsilon\}.$

T_ϵ is NOT a stopping time

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 tighter than those for gradient estimates and yet tighter than Hessian estimates.
- Algorithms can converge even with constant probability of "iteration failure."
- Interesting open problems remain - more algorithms, more assumptions on the accuracy.

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.
- A. S Berahas, L. Cao, and K. Scheinberg "Global convergence rate analysis of a generic line search algorithm with noise". *arXiv:1910.04055, 2019*,
- F. E. Curtis, K. Scheinberg, "Adaptive Stochastic Optimization". *arXiv: 2001.06699*, 2020.