Complexity analysis framework of adaptive stochastic optimization methods via martingales.

Katya Scheinberg

joint work with A. Berahas (U. Michigan), J. Blanchet (Stanford), L. Cao (Lehigh), C. Cartis (Oxford), F. Curtis (Lehigh), B. Jin (Cornell), C. Meng (Cornell), M. Menickelly (Argonne), C. Paquette (McGill) and M. Xie (Cornell)

February, 15 2021

Cornell University
Operations Research and
Information Engineering

Unconstrained Optimization

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

- We will assume throughout that f is sufficiently smooth and 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).
- How can adaptive deterministic methods be used and analyzed in nondeterministic (possibly black box) 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

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

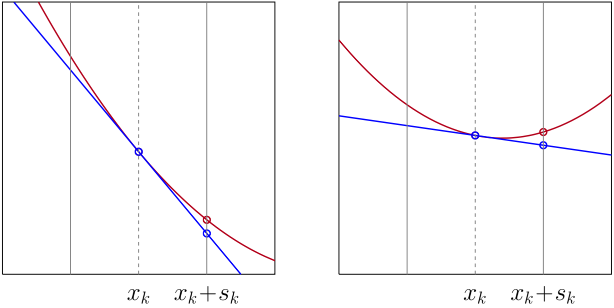


Figure: Illustration of successful (left) and unsuccessful (right) steps in a trust region method.

Why analyze adaptive methods in stochastic setting?

- For gradient descent $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ small enough step $\alpha_k \leq \frac{1}{L}$ always works.
- For inexact gradient descent $x_{k+1} = x_k - \alpha_k g_k$, $g_k \approx \nabla f(x_k)$ bound on α_k is harder to determine.
- Suppose a descent direction condition, e.g. $\|\nabla f(x_k) - g_k\| \leq \theta \|\nabla f(x_k)\|$, holds only w.p. $1 - \delta$. What kind of convergence result we can guaranatee then?
- It takes $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations until $\|\nabla f(x_k)\| \leq \epsilon$. So if for each of the first $\mathcal{O}(\frac{1}{\epsilon^2})$ iterations g_k is a descent direction, then the algorithm works!!
- Thus convergence result holds with probability $(1-\delta)^{\mathcal{O}(\frac{1}{\epsilon^2})}$.
- But what happens if the descent condition is failed even once?

First and Second order model requirements

Use a model $m_k(x_k + s) = f_k + g_k^T s + \frac{1}{2} s^T H s$.

First order model conditions

- $|f(x_k) - f_k| \leq \mathcal{O}\|s\|^2$
- $\|\nabla f(x_k) - g_k\| \leq \mathcal{O}\|s\|^1$
- $\|\nabla^2 f(x_k) - H_k\| \leq \mathcal{O}\|s\|^0$

Second order model conditions

- $|f(x_k) - f_k| \leq \mathcal{O}\|s\|^3$
- $\|\nabla f(x_k) - g_k\| \leq \mathcal{O}\|s\|^2$
- $\|\nabla^2 f(x_k) - H_k\| \leq \mathcal{O}\|s\|^1$

We consider three different cases:

- Model conditions hold deterministically - this is already known and analyzed.
- Conditions on f hold deterministically, and on g and H hold w.p. $1 - \delta$.
- Conditions on f, g and H hold w.p. $1 - \delta$.

Analysis should consider what can happens when model conditions fail to hold.

Framework for Convergence Rate Analysis, Case 1

- $\{\Phi_k\} \geq 0$ - 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 values obeying $\alpha_{k+1} = \gamma^{W_k}\alpha_k$
- T_ε, the *stopping time*, is the index of the first iterate that satisfies a desired ε-convergence criterion.

Condition 1

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

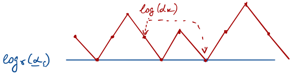
1. There exists a scalar $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that for each $k \in \mathbb{N}$, $\alpha_k \leq \underline{\alpha}_\varepsilon$ implies $W_k = 1$. Therefore, $\alpha_k \geq \underline{\alpha}_\varepsilon$ for all $k \in \mathbb{N}$.
2. There exists a nondecreasing function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

Under Condition 1

$$T_\varepsilon \leq \mathcal{O}\left(\frac{\Phi_0}{h_\varepsilon(\underline{\alpha}_\varepsilon)}\right)$$

Framework for Convergence Rate Analysis, Case 1

- $\{\Phi_k\} \geq 0$ - 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 values obeying $\alpha_{k+1} = \gamma^{W_k} \alpha_k$
- T_ε, the *stopping time*, is the index of the first iterate that satisfies a desired ε-convergence criterion.



$\Rightarrow$

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 $\alpha_{k+1} \leftarrow \min\{\gamma\alpha_k, \overline{\alpha}\}$.

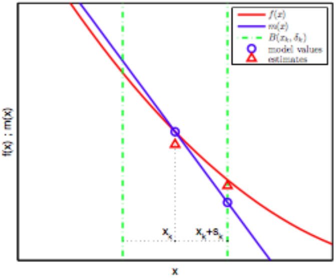
4. Unsuccessful iteration

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

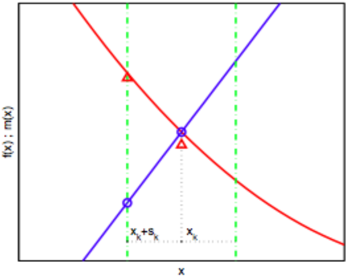
5. Next iteration

   Set $k \leftarrow k+1$.

What can happen under random models (Case 2)?



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



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

Casting the Algorithm as a Stochastic Process, Case 2

- $\{\Phi_k\} \geq 0$ - 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 values that obeying $\alpha_{k+1} = \gamma^{W_k} \alpha_k$
- T_ε, the random *stopping time*, is the index of the first iterate that satisfies a desired ε-convergence criterion.

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

Recall Condition 1

The statement 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 $h_\varepsilon : [0, \infty) \to (0, \infty)$ such that, for all $k < T_\varepsilon$, if k is successful then $\Phi_k - \Phi_{k+1} \geq h_\varepsilon(\alpha_k)$.

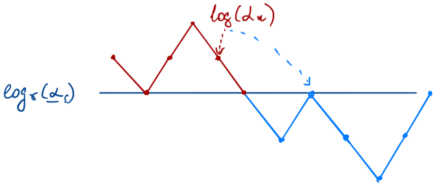
The α_k Process

Modifying Condition 1

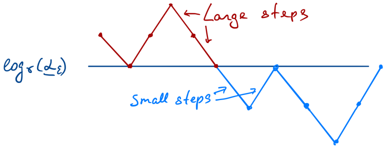
There exists a constant $\underline{\alpha}_\varepsilon \in (0,\infty)$ such that, for $k < T_\varepsilon$

$$\alpha_{k+1} \geq \gamma^{W_k}\alpha_k,$$

where $\mathbb{P}(W_k = 1 | \, \alpha_k \leq \underline{\alpha}_\varepsilon) \geq 1-\delta$.



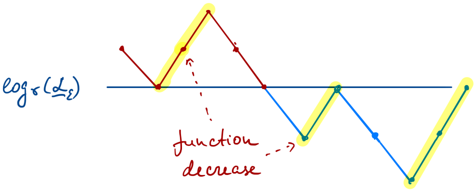
Bounding the total number of iterations



Main Ideas:

- α_k may become arbitrarily small, but it tends to increase up to $\underline{\alpha}_\varepsilon$.
- Large steps imply large function decrease, i.e. each successful iteration with accurate model and $\alpha_k \geq \underline{\alpha}_\varepsilon$ brings $h_\varepsilon(\underline{\alpha}_\varepsilon)$ improvement, so their total number is bounded.
- The number of small upward steps is bounded by the small downward steps, but downwards steps are bounded by upward steps (because of the new Condition 1).
- The number of successful iterations with accurate models and $\alpha_k \geq \underline{\alpha}_\varepsilon$ is constant fraction of the total number of iterations.

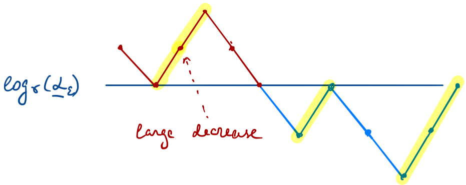
Bounding the total number of iterations



Main Ideas:

- α_k may become arbitrarily small, but it tends to increase up to $\underline{\alpha}_\varepsilon$.
- Large steps imply large function decrease, i.e. each successful iteration with accurate model and $\alpha_k \geq \underline{\alpha}_\varepsilon$ brings $h_\varepsilon(\underline{\alpha}_\varepsilon)$ improvement, so their total number is bounded.
- The number of small upward steps is bounded by the small downward steps, but downwards steps are bounded by upward steps (because of the new Condition 1).
- The number of successful iterations with accurate models and $\alpha_k \geq \underline{\alpha}_\varepsilon$ is constant fraction of the total number of iterations.

Bounding the total number of iterations



Main Ideas:

- α_k may become arbitrarily small, but it tends to increase up to $\underline{\alpha}_\varepsilon$.
- Large steps imply large function decrease, i.e. each successful iteration with accurate model and $\alpha_k \geq \underline{\alpha}_\varepsilon$ brings $h_\varepsilon(\underline{\alpha}_\varepsilon)$ improvement, so their total number is bounded.
- The number of small upward steps is bounded by the small downward steps, but downwards steps are bounded by upward steps (because of the new Condition 1).
- The number of successful iterations with accurate models and $\alpha_k \geq \underline{\alpha}_\varepsilon$ is constant fraction of the total number of iterations.

Complexity bounds

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

Condition 1

1. For all $k < T_\varepsilon$ such that $\alpha_k \leq \underline{\alpha}_\varepsilon$, $W_k = 1$ w.p.$1 - \delta$.
2. There exists a nondecreasing function $h_\varepsilon : [0, \infty) \to (0, \infty)$ such that, for all $k < T_\varepsilon$, if k is successful then $\Phi_k - \Phi_{k+1} \geq h_\varepsilon(\alpha_k)$.

Theorem

Under Condition 1,

$$\mathbb{E}[T_\varepsilon] \leq \mathcal{O}\left(\frac{1}{1-2\delta}\frac{\Phi_0}{h_\varepsilon(\underline{\alpha}_\varepsilon)}\right)$$

Moreover,

$$\mathbb{P}(T_\varepsilon \geq N) \leq e^{\left(-\frac{(\delta-\hat{\delta})^2}{2}N\right)}, \forall N \geq \mathcal{O}\left(\frac{1}{1-2\hat{\delta}}\frac{\Phi_0}{h_\varepsilon(\underline{\alpha}_\varepsilon)}\right)$$

Complexity bounds for particular cases

Line Search

For the line search algorithm with random first order models, accurate w.p $1-\delta$

- applied to nonconvex $f(x)$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\varepsilon^2}\right), \quad T_\varepsilon = \min\{k : \|\nabla f(x_k)\| \le \varepsilon\}$$

- applied to convex $f(x)$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\varepsilon}\right), \quad T_\varepsilon = \min\{k : f(x_k)-f_* \le \varepsilon\}$$

- and strongly convex $f(x)$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\log(\varepsilon)}\right), \quad T_\varepsilon = \min\{k : f(x_k)-f_* \le \varepsilon\}$$

Complexity bounds for particular cases

Trust region and Regularized Newton

- For the trust region method with random first order models, accurate w.p. $1-\delta$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\varepsilon^2}\right), \quad T_\varepsilon = \min\{k : \|\nabla f(x_k)\| \le \varepsilon\}$$

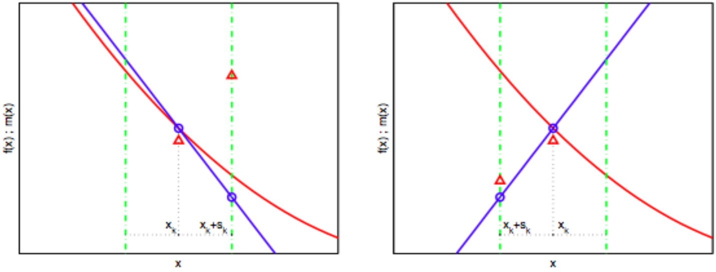
- with random second order models, accurate w.p. $1-\delta$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\varepsilon^3}\right), \; T_\varepsilon = \min\{k : \|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k)) \le \varepsilon\}$$

- For cubicly regularized Newton method with random first order models accurate w.p. $1-\delta$

$$T_\varepsilon \approx \mathcal{O}\left(\frac{1}{1-2\delta}\frac{f(x_0)-f_*}{\varepsilon^{\frac{3}{2}}}\right), \quad T_\varepsilon = \min\{k : \|\nabla f(x_k)\| \le \varepsilon\}$$

What can happen under random function estimates, Case 3



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

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

Assumptions on Stochastic Process, Case 3

- $\{\Phi_k\} \geq 0$ - 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 values that obeying $\alpha_{k+1} = \gamma^{W_k} \alpha_k$
- T_ε, the random *stopping time*, is the index of the first iterate that satisfies a desired ε-convergence criterion.

$\{\Phi_k, \alpha_k, W_k\}$ 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. $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that, for $k < T_\varepsilon$ for which $\alpha_k \leq \underline{\alpha}_\varepsilon$,
$$\alpha_{k+1} \geq \gamma^{W_k} \alpha_k, \text{ where } W_k = 1 \text{ w.p.} 1 - \delta.$$
2. There exists a nondecreasing function $h_\varepsilon : [0, \infty) \to (0, \infty)$ such that, for all $k < T_\varepsilon$, if k is successful, $\Phi_k - \Phi_{k+1} \geq h_\varepsilon(\alpha_k)$.

Assumptions on Stochastic Process, Case 3

- $\{\Phi_k\} \geq 0$ - 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 values that obeying $\alpha_{k+1} = \gamma^{W_k} \alpha_k$
- T_ε, the random *stopping time*, is the index of the first iterate that satisfies a desired ε-convergence criterion.

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

New Condition 1

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

1. $\underline{\alpha}_\varepsilon \in (0, \infty)$ such that, for $k < T_\varepsilon$ for which $\alpha_k \leq \underline{\alpha}_\varepsilon$,
$$\alpha_{k+1} \geq \gamma^{W_k} \alpha_k, \text{ where } W_k = 1 \text{ w.p.} 1 - \delta.$$
2. There exists a nondecreasing function $h(\cdot) : [0, \infty) \to (0, \infty)$ such that, until the stopping time:
$$\mathbb{E}(\Phi_{k+1} | \text{past}) \leq \Phi_k - h(\alpha_k).$$

Bounding expected stopping time

Main Idea: This is a renewal-reward process and Φ_k is a supermartingale - $\mathbb{E}[\Phi_{k+1} | \text{past}] \leq \Phi_k - h_\varepsilon(\alpha_k)$ and, thus,

- $\Phi_0 \geq \mathbb{E}[\sum_{i=0}^{T_\epsilon} h(\alpha_i)]$.
- 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 1 hold. Then

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

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)

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

Stochastic TR: Second-order convergence rate

- α_k is the trust region radius.
- $\Phi_k = \nu(f(x_k) - f_{\min}) + (1-\nu)\alpha_k^3$.
- $T_\epsilon = \inf\{k \geq 0 : \max\{\|\nabla f(x_k)\|, -\lambda_{\min}(\nabla^2 f(x_k))\} \leq \epsilon\}$.

Theorem

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

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

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)

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

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)

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

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)

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

Cubicly regularized Newton

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

T_ϵ is NOT a stopping time.Need to modify Condition 1 again.

Conclusions and Remarks

- We have a versatile framework based on bounding stoping time of a martingale which can be used to derive expected complexity bounds for adaptive stochastic methods.
- Algorithms can converge even with constant (and quite large) probability of "iteration failure."
- To do: High probability results for stochastic case.
- To do: Weaker conditions for stochastic case.
- To do: Stochastic Cubicly regularized Newton and optimal Trust Region method.

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.