The Theoretical analysis of Trust-Region Methods in Derivative-Free Optimization

Liyuan Cao 曹立元

Beijing International Center for Mathematical Research, Peking University
北京大学北京国际数学研究中心

中科院SIAM学生分会学术交流活动

April 27, 2023

Contents

Katya Scheinberg

Albert S. Berahas

DFO Problems

black-box optimization / derivative-free optimization / zeroth-order optimization

$$\min_{x \in \mathcal{X}} f(x)$$

Applications:

- simulation-based optimization
- hyperparameter tuning
- neural network adversarial attack ...

Difficulties:

- only zeroth-order information, possibly very expensive to obtain and/or noisy
- unknown structure of f: smoothness, convexity, noise level, type of constraints, etc...

DFO Algorithms

- grid search and direct/pattern search
  - simplex methods (Nelder-Mead ...)
  - directional direct-search
  - mesh adaptive direct search
  - DIviding RECTangles (DIRECT) ...
- meta-heuristics
  - simulated annealing
  - particle swarm optimization
  - genetic algorithm . . .
- derivative-based methods with derivatives estimated by finite difference
  - finite difference + gradient descent or Newton's method
  - implicit filtering
  - Nesterov's gradient free method ...
- Bayesian optimization
- evolutionary strategies (e.g. CMA-ES)
- **Powell's DFO algorithms**
- . . .

Contents

The objective function f is

- (pretty much) smooth,
- (almost) noiseless,
- very expensive to evaluate.

Algorithm: The Skeleton of Powell's DFO algorithms

Inputs: starting point x_0, starting trust-region radius δ_0, and other hyperparameters

for $k = 0, 1, 2, \ldots$ **do**

1 **Sample set management**:

Choose a sample set $\mathcal{Y}_k \subset \mathbb{R}^n$.

(reuse) Most points in $\mathcal{Y}_k$ are evaluated in the previous iterations.

Evaluate $f(y)$ for all $y \in \mathcal{Y}_k$ that has not been evaluated.

2 **Polynomial interpolation**:

Use a linear or quadratic function m_k to interpolate f on $\mathcal{Y}_k$.

3 **Trust-region method**:

Calculate $x^+ = \arg\min\{m_k(x) : \|x - x_k\| \leq \delta_k\}$ and evaluate $f(x^+)$.

Assign values to x_{k+1} and δ_{k+1} accordingly.

Powell's DFO algorithms

COBYLA Constrained Optimization BY Linear Approximation

M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In S. Gomez and J. P. Hennart, editors, *Advances in Optimization and Numerical Analysis*, pages 51–67, Dordrecht, NL, 1994. Springer.

UOBYQA Unconstrained Optimization BY Quadratic Approximation

M. J. D. Powell. UOBYQA: unconstrained optimization by quadratic approximation. *Mathematical Programming, 92:555–582, 2002.*

NEWUOA (probably) NEW Unconstrained Optimization Algorithm

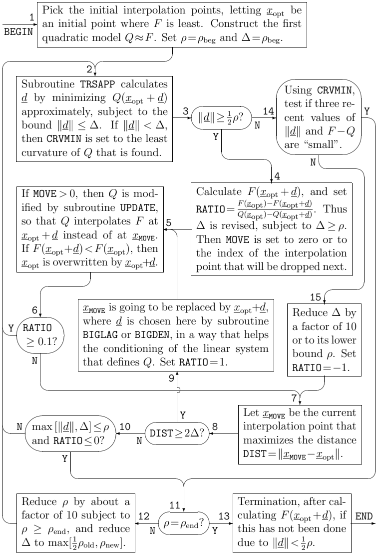
M. J. D. Powell. The NEWUOA software for unconstrained optimization without derivatives. In G. Di Pillo and M. Roma, editors, *Large-Scale Nonlinear Optimization*, volume 83 of Nonconvex Optimization and Its Applications, pages 255–297, Boston, MA, USA, 2006. Springer.

BOBYQA Bound Optimization BY Quadratic Approximation

M. J. D. Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK, 2009.

LINCOA LINearly Constrained Optimization Algorithm

Powell's DFO algorithms



An outline of the NEWUOA algorithm

DFO-TR

Algorithm: DFO-TR: a Simple Derivative-Free Trust-Region Method

Inputs: starting $x_0, \mathcal{Y}_0, \delta_0$, and $0 < \gamma_1 < 1 < \gamma_2$ and $0 < \eta_1 < \eta_2 < 1$.

for $k = 0, 1, 2, \ldots$ **do**

Sample set management:

$$\mathcal{Y}_k = \begin{cases} \mathcal{Y}_0 & \text{if } k = 0, \\ \mathcal{Y}_{k-1} \cup \{x_{k-1}^+\} & \text{if } |\mathcal{Y}_{k-1}| < (n+2)(n+1)/2, \\ \mathcal{Y}_{k-1} \setminus \arg\max\limits_{y \in \mathcal{Y}_{k-1}} \{\|y - x_k\|\} \cup \{x_{k-1}^+\} & \text{otherwise.} \end{cases}$$

Polynomial interpolation:

Build a quadratic model $m_k(x_k + s) = f_k + g_k^\intercal s + s^\intercal H_k s/2$ by solving

$$\min_{f_k, g_k, H_k} \|H_k\|_F \quad \text{s.t. } m_k(y) = f(y) \text{ for all } y \in \mathcal{Y}_k.$$

Trust-region method:

Calculate $x_k^+ = \arg\min\{m_k(x) : \|x - x_k\| \le \delta_k\}$ and evaluate $f(x_k^+)$.
Calculate $\rho_k = (f(x_k) - f(x_k^+))/(m_k(x_k) - m_k(x_k^+))$.

$$(x_{k+1}, \delta_{k+1}) = \begin{cases} (x_k, \gamma_1 \delta_k) & \text{if } \rho_k < \eta_1, \\ (x^+, \delta_k) & \text{if } \eta_1 \le \rho_k < \eta_2, \\ (x^+, \gamma_2 \delta_k) & \text{if } \rho_k \ge \eta_2. \end{cases}$$

Contents

The Noisy Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

f follows common assumptions

$$\begin{aligned} f(x) \geq f_* \text{ for all } x \in \mathbb{R}^n, &\text{ and} \\ \|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\| &\text{ for all } (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \\ \text{or } \|\nabla^2 f(x) - \nabla^2 f(y)\| \leq L_2 \|x - y\| &\text{ for all } (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \end{aligned}$$

but we only have access to

$$\left\{ \begin{array}{l} f_k = f(x_k) + e(x_k) \\ g_k = \nabla m_k(x_k) \\ H_k = \nabla^2 m_k(x_k) \end{array} \right. \quad \text{instead of} \quad \left\{ \begin{array}{l} f(x) \\ \nabla f(x) \\ \nabla^2 f(x). \end{array} \right.$$

Important Concepts

Definition (fully linear model)

A function m_k is a $(\kappa_{eg}, \kappa_{ef})$-fully linear model of f on $B(x_k, \delta_k)$, if for every $x \in B(x_k, \delta_k)$,

$$\|\nabla f(x) - \nabla m_k(x)\| \leq \kappa_{eg}\delta_k,$$
$$|f(x) - m_k(x)| \leq \kappa_{ef}\delta_k^2.$$

Definition (fully quadratic model)

A function m_k is a $(\kappa_{eh}, \kappa_{eg}, \kappa_{ef})$-fully quadratic model of f on $B(x_k, \delta_k)$, if for every $x \in B(x_k, \delta_k)$,

$$\|\nabla^2 f(x) - \nabla^2 m_k(x)\| \leq \kappa_{eh}\delta_k,$$
$$\|\nabla f(x) - \nabla m_k(x)\| \leq \kappa_{eg}\delta_k^2,$$
$$|f(x) - m_k(x)| \leq \kappa_{ef}\delta_k^3.$$

Definition (probabilistically sufficiently accurate model)

A sequence of random models $\{M_k\}$ is p-probabilistically fully linear/quadratic for a corresponding sequence $\{B(X_k, \Delta_k)\}$ if it satisfies the following submartingale condition

$$\mathbb{P}\left\{M_k \text{ is a fully linear/quadratic model of } f \text{ on } B(X_k, \Delta_k) | \mathcal{F}_k\right\} \geq p.$$

A.S. Bandeira, K. Scheinberg, L.N. Vicente. Computation of sparse low degree interpolating polynomials and their application to derivative-free optimization. *Mathematical programming*. 2012 Aug;134(1):223-57.

Applying compressive sensing to trust-region DFO:

$$\min_{f_k, g_k, H_k} \|\mathrm{vec}(H_k)\|_1 \quad \text{s.t.} \; \sum_{i=1}^{p} (m_k(y_i) - f(y_i))^2 \leq \eta.$$

Theorem

Assume f is twice differentiable, and its Hessian is Lipschitz continuous and h-sparse. Given x, δ, and a set of p random points $\mathcal{Y} = \{y_1, \ldots, y_p\}$ chosen with respect to the uniform measure in $B_\infty(x;\delta)$ with

$$\frac{p}{\log p} \geq 9c_1(h+n+1)(\log(h+n+1))^2 \log q,$$

the solution to the above problem provides a $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$-fully quadratic model of f on $B_\infty(x;\delta)$ with probability larger than $1 - n^{-c_2 \log p}$. The constants c_1 and c_2 are universal and $(\kappa_{ef}, \kappa_{eg}, \kappa_{eh})$ do not depend on x or δ.

On the Convergence of Stochastic TR

$\mathcal{S}^{(0)} = [0, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = (0, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- (c) Afonso S Bandeira, Katya Scheinberg, and Luis Nunes Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.
- (h) Serge Gratton, Clement W Royer, Luis N Vicente, and Zaikun Zhang. Complexity and global rates of trust-region methods base on probabilistic models. *IMA Journal of Numerical Analysis*, 38(3):1579-1597, 2018.

$\mathcal{S}^{(j)} = (0, \infty) \times [0, \bar{p}_j]$ for sufficiently large $\bar{p}_j$, $j = 0, 1$:

- (c) Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.

$\mathcal{S}^{(j)} = (0, \infty) \times [0, \bar{p}_j]$ for sufficiently large $\bar{p}_j$, $j = 0, 1, 2$ and $\mathbb{E}_{\xi_0}|f_k - f(x_k)| \le C_0$:

- (e) Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

$\mathcal{S}^{(0)} = [\epsilon_f, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = [\epsilon_g, \infty) \times [0, 1]$:

- Shigeng Sun and Jorge Nocedal. A trust region method for the optimization of noisy functions. *arXiv preprint arXiv:2201.00973*, 2022.

$\mathcal{S}^{(0)} = \{(e, p) : e \ge \epsilon_f, p \le 1 - \exp(a(\epsilon_f - e))\}$ for some $a > 0$ and $\epsilon_f \ge 0$ and $\mathcal{S}^{(1)} = (\epsilon_g, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- (h) Liyuan Cao, Albert S. Berahas, and Katya Scheinberg. First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. arXiv preprint arXiv:2205.03667. 2022 May 7.

The Trust-Region Method

Algorithm: Modified First-Order Trust Region Algorithm

Inputs: Starting point x_0, initial trust region radius δ_0, tolerance parameter r, and hyperparameters $\eta_1 > 0, \eta_2 > 0$, $\gamma \in (0,1)$ for controlling the trust region radius.

for $k = 0, 1, 2, \ldots$ **do**

Build a quadratic model $m_k(x_k + s) = f(x_k) + \langle g_k, s\rangle + 0.5 \langle H_k s, s\rangle$

Compute s_k by approximately minimizing m_k in $B(x_k, \delta_k)$ so that it satisfies the *Cauchy decrease condition*

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{1}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \delta_k\right\}.$$

Compute

$$\rho_k = \frac{f_k - f_k^+ + r}{m_k(x_k) - m_k(x_k + s_k)}$$

if $\rho_k \geq \eta_1$ **then**

Set $x_{k+1} = x_k + s_k$ and

$$\delta_{k+1} = \begin{cases} \gamma^{-1}\delta_k & \text{if } \|g_k\| \geq \eta_2 \delta_k \\ \gamma \delta_k, & \text{if } \|g_k\| < \eta_2 \delta_k \end{cases}$$

else

Set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma \delta_k$

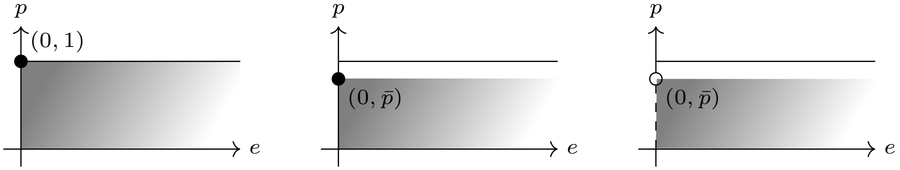
Stochastic Oracles

Let $\varphi^{(j)}\left(x_k, \xi_k^{(j)}, \mathcal{S}_k^{(j)}\right)$ be the jth-order oracle that returns an estimate of $\nabla^j f(x_k)$ such that

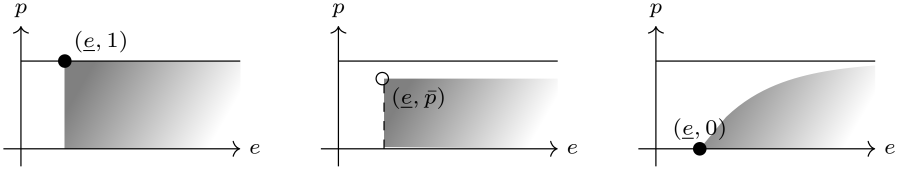
$$\mathbb{P}_{\xi_k^{(j)}}\left\{\left\|\varphi^{(j)}\left(x_k, \xi_k^{(j)}, \mathcal{S}^{(j)}\right) - \nabla^j f(x_k)\right\| \leq e \,\middle|\, \mathcal{F}_k\right\} \geq p$$

for all

$$(e, p) \in \mathcal{S}_k^{(j)} \subset [0, \infty) \times [0, 1].$$



(a) $\mathcal{S}_k^{(j)} \ni (0, 1)$ **(b)** $\mathcal{S}_k^{(j)} = [0, +\infty) \times [0, \bar{p}]$ **(c)** $\mathcal{S}_k^{(j)} = (0, +\infty) \times [0, \bar{p}]$



(d) $\mathcal{S}_k^{(j)} = [\underline{e}, +\infty) \times [0, 1]$ **(e)** $\mathcal{S}_k^{(j)} = (\underline{e}, +\infty) \times [0, \bar{p}]$ **(f)** $\mathcal{S}_k^{(j)} = \{(e, p) : e \geq \underline{e}, p \leq 1 - \exp(a(\underline{e} - e))\}$

Stochastic Oracles

Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 2019.

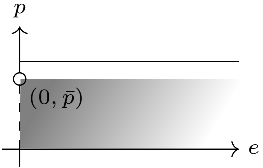
Assumption

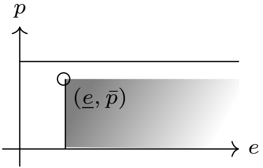
1. *The model Hessians satisfy* $\|H_k\|_2 \le \kappa_{bhm}$ *for some* $\kappa_{bhm} \ge 1$ *for all* k *deterministically.*
2. *The sequence of random models* $\{M_k\}$ *is* p_m*-probabilistically* $(\kappa_{ef}, \kappa_{eg})$*-fully linear for sufficiently large* p_m*, i.e.,*

$$\mathbb{P}\left(\begin{matrix}\|\nabla f(X_k) - \nabla M_k(X_k)\| \le \kappa_{eg}\Delta_k \text{ and} \\ |f(x) - M_k(x)| \le \kappa_{ef}\Delta_k^2 \ \forall x \in B(X_k, \Delta_k)\end{matrix}\middle| \mathcal{F}_k\right) \ge p_m.$$

3. *The sequence of random estimates* $\{(F_k, F_k^+)\}$ *is* p_0*-probabilistically* ϵ_f*-accurate for sufficiently large* p_0 *and sufficiently small* ϵ_f*, i.e.,*

$$\mathbb{P}\left(\begin{matrix}|f(X_k) - F_k| \le \epsilon_f \text{ and} \\ |f(X_k^+) - F_k^+| \le \epsilon_f\end{matrix}\middle| \mathcal{F}_k\right) \ge p_0.$$





Oracle Assumptions

Liyuan Cao, Albert S. Berahas, and Katya Scheinberg. First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. arXiv preprint arXiv:2205.03667. 2022 May 7.

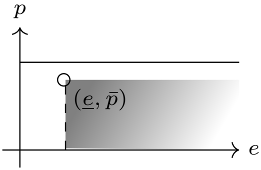
Assumption

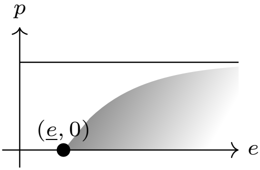
- *The model Hessians satisfy* $\|H_k\|_2 \le \kappa_{bhm}$ *for some* $\kappa_{bhm} \ge 1$ *for all* k *deterministically.*
- *The sequence of random gradients* $\{\nabla M_k\}$ *is* p_1*-probabilistically* $(\kappa_{eg}, \epsilon_g)$*-sufficiently accurate for sufficiently large* p_1*, i.e.,*

$$\mathbb{P}\left\{\|\nabla M_k(X_k) - \nabla f(X_k)\| \le \kappa_{\mathrm{eg}}\Delta_k + \epsilon_g | \mathcal{F}_k\right\} \ge p_1.$$

- *The sequence of random estimates* $\{(F_k, F_k^+)\}$ *is* (ϵ_f, a)*-subexponentially distributed, i.e.,*

$$\left.\begin{array}{l} \mathbb{P}\left\{|F_k - f(X_k)| \le e | \mathcal{F}_k\right\} \\ \mathbb{P}\left\{|F_k^+ - f(X_k^+)| \le e \middle| \mathcal{F}_k\right\} \end{array}\right\} \ge \exp(a(\epsilon_f - e)).$$





Oracle Assumptions

Liyuan Cao, Albert S. Berahas, and Katya Scheinberg. First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. arXiv preprint arXiv:2205.03667. 2022 May 7.

Assumption

- *The model Hessians satisfy* $\|H_k\|_2 \le \kappa_{bhm}$ *for some* $\kappa_{bhm} \ge 1$ *for all* k *deterministically.*
- *The sequence of random gradients* $\{\nabla M_k\}$ *is* p_1*-probabilistically* $(\kappa_{eg}, \epsilon_g)$*-sufficiently accurate for sufficiently large* p_1*, i.e.,*

$$\mathbb{P}\left\{\|\nabla M_k(X_k) - \nabla f(X_k)\| \le \kappa_{\mathrm{eg}}\Delta_k + \epsilon_g | \mathcal{F}_k\right\} \ge p_1.$$

- *The sequence of random estimates* $\{(F_k, F_k^+)\}$ *is* (ϵ_f, a)*-subexponentially distributed, i.e.,*

$$\left.\begin{array}{l} \mathbb{P}\left\{|F_k - f(X_k)| \le e | \mathcal{F}_k\right\} \\ \mathbb{P}\left\{|F_k^+ - f(X_k^+)| \le e | \mathcal{F}_k\right\} \end{array}\right\} \ge \exp(a(\epsilon_f - e)).$$

Lemma

If $\|g_k - \nabla f(x_k)\| \le \kappa_{\mathrm{eg}}\delta_k + \epsilon_g$ *holds, then*

$$|m_k(x) - f(x)| \le (L_1 + \kappa_{\mathrm{bhm}} + 2\kappa_{\mathrm{eg}})\delta_k^2/2 + \epsilon_g\delta_k$$

for all $x \in B(x_k, \delta_k)$.

Types of Analysis Results

ⓒ convergence

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| \leq \epsilon$$

ⓔ expected complexity

$$\mathbb{E} \min\{k : \|\nabla f(X_k)\| \leq \epsilon\} = \mathcal{O}(1/\epsilon^2)$$

ⓗ high probability convergence

$$\mathbb{P}\left\{\min\{\|\nabla f(X_k)\| : \ 0 \leq k \leq T-1\} < \epsilon\right\}$$
$$\geq \text{a function of } T \text{ the converges to 1 as } T \text{ increase}$$

for $T \geq \mathcal{O}(1/\epsilon^2)$ and some sufficiently large ϵ.

$\mathcal{S}^{(0)} = [0, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = (0, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- Afonso S Bandeira, Katya Scheinberg, and Luis Nunes Vicente. Convergence of trust-region methods based on probabilistic models. *SIAM Journal on Optimization*, 24(3):1238–1264, 2014.
- Serge Gratton, Clement W Royer, Luis N Vicente, and Zaikun Zhang. Complexity and global rates of trust-region methods base on probabilistic models. *IMA Journal of Numerical Analysis*, 38(3):1579-1597, 2018.

$\mathcal{S}^{(j)} = (0, \infty) \times [0, \bar{p}_j]$ for sufficiently large $\bar{p}_j$, $j = 0, 1$:

- Ruobing Chen, Matt Menickelly, and Katya Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169(2):447–487, 2018.

$\mathcal{S}^{(j)} = (0, \infty) \times [0, \bar{p}_j]$ for sufficiently large $\bar{p}_j$, $j = 0, 1, 2$ and $\mathbb{E}_{\xi_0}|f_k - f(x_k)| \le C_0$:

- Jose Blanchet, Coralia Cartis, Matt Menickelly, and Katya Scheinberg. Convergence rate analysis of a stochastic trust-region method via supermartingales. *INFORMS Journal on Optimization*, 1(2):92–119, 2019.

$\mathcal{S}^{(0)} = [\epsilon_f, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = [\epsilon_g, \infty) \times [0, 1]$:

- Shigeng Sun and Jorge Nocedal. A trust region method for the optimization of noisy functions. *arXiv preprint arXiv:2201.00973*, 2022.

$\mathcal{S}^{(0)} = \{(e, p) : e \ge \epsilon_f, p \le 1 - \exp(a(\epsilon_f - e))\}$ for some $a > 0$ and $\epsilon_f \ge 0$ and
$\mathcal{S}^{(1)} = (\epsilon_g, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- Liyuan Cao, Albert S. Berahas, and Katya Scheinberg. First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. arXiv preprint arXiv:2205.03667. 2022 May 7.

$\mathcal{S}^{(0)} = [0, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = (0, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- Coralia Cartis and Katya Scheinberg. Global convergence rate analysis of unconstrained optimization methods based on probabilistic models. *Mathematical Programming*, 169(2):337–375, 2018.

$\mathcal{S}^{(j)} = (0, \infty) \times [0, \bar{p}_j]$ for sufficiently large $\bar{p}_j$, $j = 0, 1$ and $\mathbb{E}_{\xi_0}|f_k - f(x_k)| \le C_0$:

- Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

$\mathcal{S}^{(0)} = [\epsilon_f, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = [\epsilon_g, \infty) \times [0, 1]$:

- Albert S Berahas, Richard H Byrd, and Jorge Nocedal. Derivative-free optimization of noisy functions via quasi-newton methods. *SIAM Journal on Optimization*, 29(2):965–993, 2019.

$\mathcal{S}^{(0)} = [\epsilon_f, \infty) \times [0, 1]$ and $\mathcal{S}^{(1)} = (0, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- Albert S Berahas, Liyuan Cao, and Katya Scheinberg. Global convergence rate analysis of a generic line search algorithm with noise. *SIAM Journal on Optimization*, 31(2):1489–1518, 2021.

subexponential zeroth-order noise and $\epsilon_f \ge 0$ and $\mathcal{S}^{(1)} = (0, \infty) \times [0, \bar{p}_1]$ for sufficiently large $\bar{p}_1$:

- Jin, Billy, Katya Scheinberg, and Miaolan Xie. High probability complexity bounds for line search based on stochastic oracles. *Advances in Neural Information Processing Systems*, 34, 2021.

The Noisy Problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

f follows common assumptions

$$\begin{aligned}
&f(x) \geq f_* \text{ for all } x \in \mathbb{R}^n, \\
&\|\nabla f(x) - \nabla f(y)\| \leq L_1 \|x - y\| \text{ for all } (x, y) \in \mathbb{R}^n \times \mathbb{R}^n,
\end{aligned}$$

but we only have access to

$$\begin{cases} f_k = f(x_k) + e(x_k) \\ g_k = \nabla m_k(x_k) \\ H_k = \nabla^2 m_k(x_k) \end{cases} \quad \text{instead of} \quad \begin{cases} f(x) \\ \nabla f(x) \\ \nabla^2 f(x). \end{cases}$$

Algorithm: Modified First-Order Trust Region Algorithm

Inputs: Starting point x_0, initial trust region radius δ_0, tolerance parameter r, and hyperparameters $\eta_1 > 0, \eta_2 > 0$, $\gamma \in (0,1)$ for controlling the trust region radius.

for $k = 0, 1, 2, \ldots$ **do**

Build a quadratic model $m_k(x_k + s) = f(x_k) + \langle g_k, s\rangle + 0.5\,\langle H_k s, s\rangle$

Compute s_k by approximately minimizing m_k in $B(x_k, \delta_k)$ so that it satisfies the *Cauchy decrease condition*

$$m_k(x_k) - m_k(x_k + s_k) \geq \frac{1}{2}\|g_k\| \min\left\{\frac{\|g_k\|}{\|H_k\|}, \delta_k\right\}.$$

Compute

$$\rho_k = \frac{f_k - f_k^+ + r}{m_k(x_k) - m_k(x_k + s_k)}$$

if $\rho_k \geq \eta_1$ **then**

Set $x_{k+1} = x_k + s_k$ and

$$\delta_{k+1} = \begin{cases} \gamma^{-1}\delta_k & \text{if } \|g_k\| \geq \eta_2\delta_k \\ \gamma\delta_k, & \text{if } \|g_k\| < \eta_2\delta_k \end{cases}$$

else

Set $x_{k+1} = x_k$ and $\delta_{k+1} = \gamma\delta_k$

random variables:	X_k	X_k^+	$\mathcal{E}_k$	$\mathcal{E}_k^+$	
realizations:	x_k	$x_k + s_k$	$\lvert f_k - f(x_k)\rvert$	$\lvert f_k^+ - f(x_k + s_k)\rvert$	
random variables:	M_k	∇M_k	$\nabla^2 M_k$	Δ_k	ρ_k
realizations:	m_k	g_k	H_k	δ_k	ρ_k

Define

$$I_k = \mathbb{1}\{\|\nabla M_k - \nabla f(X_k)\| \leq \kappa_{\text{eg}}\Delta_k + \epsilon_g\} \qquad \text{gradient sufficiently accurate}$$

$$J_k = \mathbb{1}\{\mathcal{E}_k + \mathcal{E}_k^+ \leq r\} \qquad \text{zeroth-order noise compensated}$$

$$\Lambda_k = \mathbb{1}\{\Delta_k > \bar{\Delta}\} \qquad \text{large TR radius}$$

$$\Theta_k = \mathbb{1}\{\rho_k \geq \eta_1 \text{ and } \|\nabla M_k\| \geq \eta_2 \Delta_k\} \qquad \text{successful step}$$

$$\Theta_k' = \mathbb{1}\{\rho_k \geq \eta_1\} \qquad \text{accepted step}$$

where $\bar{\Delta} = C_1 \min_{0 \leq k \leq T-1} \|\nabla f(X_k)\| - C_2 \epsilon_g$.

Classification of Iterations

<table>
<tr><th></th><th colspan="3">$I_k = 1, J_k = 1$</th><th colspan="3">$I_k = 1, J_k = 0$</th><th colspan="3">$I_k = 0, J_k = 1$</th><th colspan="3">$I_k = 0, J_k = 0$</th></tr>
<tr><th></th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th></tr>
<tr><td>$\Delta_k \in (\bar{\Delta}, \infty)$</td><td>1</td><td>4</td><td>5</td><td>6</td><td>9</td><td>11</td><td>13</td><td>16</td><td>18</td><td>20</td><td>23</td><td>25</td></tr>
<tr><td>$\Delta_k \in (\gamma\bar{\Delta}, \bar{\Delta}]$</td><td>2</td><td colspan="2" rowspan="2"></td><td>7</td><td rowspan="2">10</td><td rowspan="2">12</td><td>14</td><td rowspan="2">17</td><td rowspan="2">19</td><td>21</td><td rowspan="2">24</td><td rowspan="2">26</td></tr>
<tr><td>$\Delta_k \in (0, \gamma\bar{\Delta}]$</td><td>3</td><td>8</td><td>15</td><td>22</td></tr>
</table>

Lemma (sufficient condition for successful step)

If $I_k J_k = 1$ *and* $\Lambda_k = 0$ *then* $\Theta_k = 1$.

Lemma (progress made in each iteration)

Let $h(\delta) = C_3\delta^2$. Then we have

$$f(X_k) - f(X_{k+1}) \geq \begin{cases} h(\Delta_k) - \mathcal{E}_k - \mathcal{E}_k^+ - r & \text{if } \Theta_k = 1 \text{ (successful)} \\ -\mathcal{E}_k - \mathcal{E}_k^+ - r & \text{if } \Theta'_k = 1 \text{ (accepted)} \\ 0 & \text{if } \Theta'_k = 0 \text{ (rejected).} \end{cases}$$

Lemma (total progress)

$$h(\gamma\bar{\Delta}) \sum_{k=0}^{T-1} \Theta_k \Lambda_k \leq f(x_0) - f_* + \sum_{k=0}^{T-1} \Theta'_k \left(\mathcal{E}_k + \mathcal{E}_k^+ + r\right).$$

Lemma of Total Progress

Lemma (total loss)

For any $t \geq 0$,

$$\mathbb{P}\left\{\sum_{k=0}^{T-1}\left(\mathcal{E}_k+\mathcal{E}_k^{+}+r\right) \geq T(4/a+2\epsilon_f+r)+t\right\} \leq \exp\left(-\frac{a}{4}t\right).$$

Let $t = rT$.

Lemma (total progress)

$$h(\gamma\bar{\Delta})\sum_{k=0}^{T-1}\Theta_k\Lambda_k \leq f(x_0)-f_*+\sum_{k=0}^{T-1}\Theta_k'\left(\mathcal{E}_k+\mathcal{E}_k^{+}+r\right)$$
$$< f(x_0)-f_*+T(4/a+2\epsilon_f+2r)$$

with probability at least $1-\exp\left(-\frac{ar}{4}T\right)$.

$$h(\gamma\bar{\Delta}) = \gamma^2 C_3\left(C_1\min_{0\leq k\leq T-1}\|\nabla f(X_k)\| - C_2\epsilon_g\right)^2$$

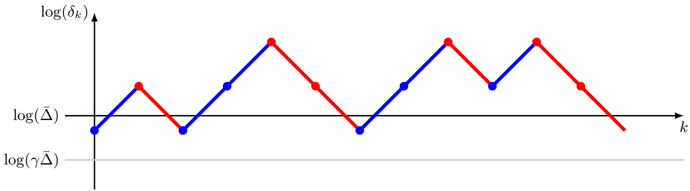
<table>
<tr><th></th><th colspan="3">$I_k = 1, J_k = 1$</th><th colspan="3">$I_k = 1, J_k = 0$</th><th colspan="3">$I_k = 0, J_k = 1$</th><th colspan="3">$I_k = 0, J_k = 0$</th></tr>
<tr><th></th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th></tr>
<tr><td>$\Delta_k \in (\bar{\Delta}, \infty)$</td><td>1</td><td>4</td><td>5</td><td>6</td><td>9</td><td>11</td><td>13</td><td>16</td><td>18</td><td>20</td><td>23</td><td>25</td></tr>
<tr><td>$\Delta_k \in (\gamma\bar{\Delta}, \bar{\Delta}]$</td><td>2</td><td rowspan="2" colspan="2"></td><td>7</td><td rowspan="2">10</td><td rowspan="2">12</td><td>14</td><td rowspan="2">17</td><td rowspan="2">19</td><td>21</td><td rowspan="2">24</td><td rowspan="2">26</td></tr>
<tr><td>$\Delta_k \in (0, \gamma\bar{\Delta}]$</td><td>3</td><td>8</td><td>15</td><td>22</td></tr>
</table>

Lemma (**total progress**)

$$\mathbb{P}\left\{ h(\gamma\bar{\Delta}) \sum_{k=0}^{T-1} \Theta_k \Lambda_k \leq f(x_0) - f_* + T(4/a + 2\epsilon_f + 2r) \right\} \geq 1 - \exp\left(-\frac{ar}{4} T \right).$$

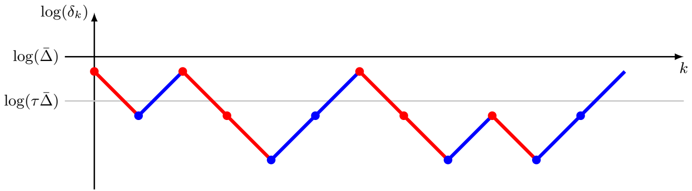
Ups and Downs of the Radius

	$I_k=1, J_k=1$			$I_k=1, J_k=0$			$I_k=0, J_k=1$			$I_k=0, J_k=0$		
	★	✓	✗	★	✓	✗	★	✓	✗	★	✓	✗
$\Delta_k \in (\bar{\Delta}, \infty)$	1	4	5	6	9	11	13	16	18	20	23	25
$\Delta_k \in (\gamma\bar{\Delta}, \bar{\Delta}]$	2			7			14			21		
$\Delta_k \in (0, \gamma\bar{\Delta}]$	3			8	10	12	15	17	19	22	24	26

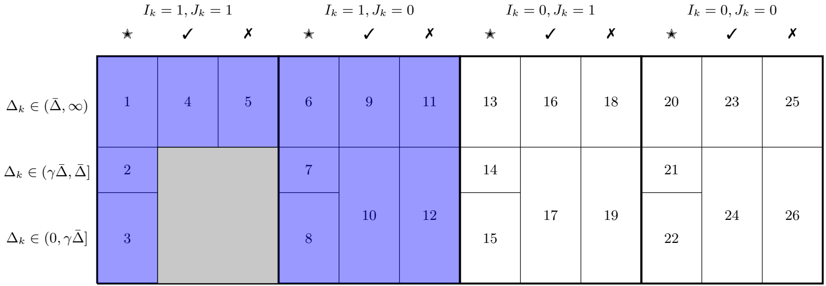


$$\color{red}{\sum_{k=0}^{T-1}(1-\Theta_k)\Lambda_k} < \color{blue}{\sum_{k=0}^{T-1}\Theta_k\Lambda_k} + \min\left\{\log_\gamma\left(\frac{\delta_0}{\bar{\Delta}}\right), 0\right\} + 1$$

	$I_k = 1, J_k = 1$			$I_k = 1, J_k = 0$			$I_k = 0, J_k = 1$			$I_k = 0, J_k = 0$		
	★	✓	✗	★	✓	✗	★	✓	✗	★	✓	✗
$\Delta_k \in (\bar{\Delta}, \infty)$	1	4	5	6	9	11	13	16	18	20	23	25
$\Delta_k \in (\gamma\bar{\Delta}, \bar{\Delta}]$	2			7	10	12	14	17	19	21	24	26
$\Delta_k \in (0, \gamma\bar{\Delta}]$	3			8			15			22		



$$\sum_{k=0}^{T-1} \Theta_k (1 - \Lambda_k) < \sum_{k=0}^{T-1} (1 - \Theta_k)(1 - \Lambda_k) + \min\left\{ \log_\gamma \left(\frac{\bar{\Delta}}{\delta_0} \right), 0 \right\} + 1$$



Lemma

Assume $\mathbb{P}\{I_k = 1 \mid \mathcal{F}_k\} \geq p_1$ *holds. By Azuma-Hoeffding inequality, for any positive integer* T *and any* $\hat{p}_1 \in [0, p_1]$ *we have*

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} I_k > \hat{p}_1 T\right\} \geq 1 - \exp\left(-\frac{(1 - \hat{p}_1/p_1)^2}{2} T\right).$$

Iterations with Sufficiently Accurate Function Evaluation

<table>
<tr><th></th><th colspan="3">$I_k = 1, J_k = 1$</th><th colspan="3">$I_k = 1, J_k = 0$</th><th colspan="3">$I_k = 0, J_k = 1$</th><th colspan="3">$I_k = 0, J_k = 0$</th></tr>
<tr><th></th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th><th>★</th><th>✓</th><th>✗</th></tr>
<tr><td>$\Delta_k \in (\bar{\Delta}, \infty)$</td><td>1</td><td>4</td><td>5</td><td>6</td><td>9</td><td>11</td><td>13</td><td>16</td><td>18</td><td>20</td><td>23</td><td>25</td></tr>
<tr><td>$\Delta_k \in (\gamma\bar{\Delta}, \bar{\Delta}]$</td><td>2</td><td colspan="2" rowspan="2"></td><td>7</td><td rowspan="2">10</td><td rowspan="2">12</td><td>14</td><td rowspan="2">17</td><td rowspan="2">19</td><td>21</td><td rowspan="2">24</td><td rowspan="2">26</td></tr>
<tr><td>$\Delta_k \in (0, \gamma\bar{\Delta}]$</td><td>3</td><td>8</td><td>15</td><td>22</td></tr>
</table>

Lemma

Assume both $\mathbb{P}\{\mathcal{E}_k > t\}$ *and* $\mathbb{P}\{\mathcal{E}_k^+ > t\}$ *are* $\leq \exp(a(\epsilon_f - t))$. *Let* $p_0 = 1 - 2\exp(a[\epsilon_f - r/2])$. *For any positive integer* T *and any* $\hat{p}_0 \in [0, p_0]$, *we have*

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} J_k > \hat{p}_0 T\right\} \geq 1 - \exp\left(-\frac{(1 - \hat{p}_0/p_0)^2}{2} T\right).$$

Analysis

$$\sum_{k=0}^{T-1}(1-\Theta_k)\Lambda_k < \sum_{k=0}^{T-1}\Theta_k\Lambda_k + \min\left\{\log_\gamma\left(\frac{\delta_0}{\bar\Delta}\right), 0\right\} + 1$$

$$\sum_{k=0}^{T-1}\Theta_k(1-\Lambda_k) < \sum_{k=0}^{T-1}(1-\Theta_k)(1-\Lambda_k) + \min\left\{\log_\gamma\left(\frac{\bar\Delta}{\delta_0}\right), 0\right\} + 1$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} I_k > \hat p_1 T\right\} \geq 1 - \exp\left(-\frac{(1-\hat p_1/p_1)^2}{2}T\right)$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1} J_k > \hat p_0 T\right\} \geq 1 - \exp\left(-\frac{(1-\hat p_0/p_0)^2}{2}T\right)$$

$$\Downarrow$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1}\Theta_k\Lambda_k > \left(\hat p_0 + \hat p_1 - \frac{3}{2}\right)T - \frac{1}{2}\left|\log_\gamma\frac{\bar\Delta}{\delta_0}\right| - \frac{1}{2}\right\}$$

$$\geq 1 - \exp\left(-\frac{(1-\hat p_1/p_1)^2}{2}T\right) - \exp\left(-\frac{(1-\hat p_0/p_0)^2}{2}T\right)$$

Main Result

$$\mathbb{P}\left\{h(\gamma\bar{\Delta})\sum_{k=0}^{T-1}\Theta_k\Lambda_k \le f(x_0) - f_* + T(4/a + 2\epsilon_f + 2r)\right\} \ge 1 - \exp\left(-\frac{ar}{4}T\right).$$

$$\mathbb{P}\left\{\sum_{k=0}^{T-1}\Theta_k\Lambda_k > \left(\hat{p}_0 + \hat{p}_1 - \frac{3}{2}\right)T - \frac{1}{2}\left|\log_\gamma \frac{\bar{\Delta}}{\delta_0}\right| - \frac{1}{2}\right\}$$
$$\ge 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right) - \exp\left(-\frac{(1-\hat{p}_0/p_0)^2}{2}T\right)$$

Lemma

When the modified 1st-order TR method is applied to L_1-Lipschitz smooth functions, it holds that

$$\mathbb{P}\left\{h(\gamma\bar{\Delta})\left[\left(\hat{p}_0 + \hat{p}_1 - \frac{3}{2}\right)T - \frac{1}{2}\left|\log_\gamma \frac{\bar{\Delta}}{\delta_0}\right| - \frac{1}{2}\right] < f(x_0) - f_* + (2\epsilon_f + 4/a + 2r)T\right\}$$
$$\ge 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right) - \exp\left(-\frac{(1-\hat{p}_0/p_0)^2}{2}T\right) - \exp\left(-\frac{ar}{4}T\right).$$

Main Result

Theorem

Let assumptions hold. Given any $\epsilon > \sqrt{\frac{4\epsilon_f + 8/a + 2r}{C_3\gamma^2 C_1^2(2p_0+2p_1-3)}} + \frac{C_2}{C_1}\epsilon_g$*, we have*

$$\mathbb{P}\left\{\min\{\|\nabla f(X_k)\| : \; 0 \le k \le T-1\} \le \epsilon\right\} \ge 1 - \exp\left(-\frac{(1-\hat{p}_1/p_1)^2}{2}T\right) - \exp\left(-\frac{(1-\hat{p}_0/p_0)^2}{2}T\right) - \exp\left(-\frac{ar}{4}T\right)$$

for any $\hat{p}_0$ *and* $\hat{p}_1$ *such that* $\hat{p}_0 + \hat{p}_1 \in \left(\frac{3}{2} + \frac{2\epsilon_f + 4/a + r}{C_3\gamma^2(C_1\epsilon - C_2\epsilon_g)^2}, p_0 + p_1\right]$*, any* $t \ge 0$*, and any*

$$T \ge \left(\hat{p}_0 + \hat{p}_1 - \frac{3}{2} - \frac{2\epsilon_f + 4/a + 2r}{C_3\gamma^2(C_1\epsilon - C_2\epsilon_g)^2}\right)^{-1} \left[\frac{f(x_0) - f_*}{C_3\gamma^2(C_1\epsilon - C_2\epsilon_g)^2} + \frac{1}{2}\left|\log_\gamma \frac{C_1\epsilon - C_2\epsilon_g}{\delta_0}\right| + \frac{1}{2}\right] = \bar{\mathcal{O}}(\epsilon^{-2}).$$

Other Results

- Analyses under bounded noise assumption.
- Second-order TR method and analysis.
- Numerically testing the strength of the theoretical results.
- Experimenting with different values for r.