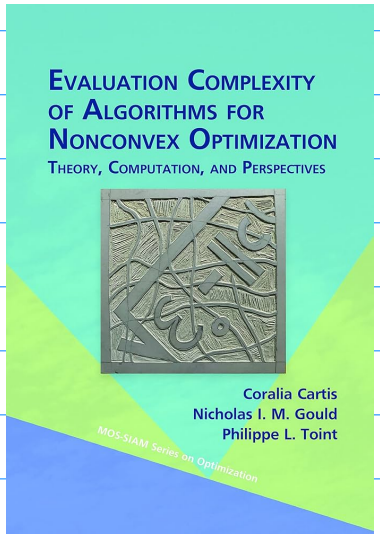
Complexity in continuous optimization (3/6)

February 12, 2025

Today: Second-order methods and guarantees

References



Math. Program., Ser. A 108, 177–205 (2006)
Digital Object Identifier (DOI) 10.1007/s10107-006-0706-8

Yurii Nesterov · B.T. Polyak

Cubic regularization of Newton method and its global performance*

Math. Program., Ser. A (2017) 162:1–32
DOI 10.1007/s10107-016-1026-2
FULL LENGTH PAPER

A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization

Frank E. Curtis[1] · Daniel P. Robinson[2] · Mohammadreza Samadi[1]

SIAM J. OPTIM.
Vol. 31, No. 1, pp. 518–544

TRUST-REGION NEWTON-CG WITH STRONG SECOND-ORDER COMPLEXITY GUARANTEES FOR NONCONVEX OPTIMIZATION*

FRANK E. CURTIS†, DANIEL P. ROBINSON†, CLÉMENT W. ROYER‡, AND STEPHEN J. WRIGHT§

(1) Second order

Problem: (P) $\min_{x \in \mathbb{R}^n} f(x)$ $\quad f : \mathbb{R}^n \to \mathbb{R}$, C^2

$\forall x \in \mathbb{R}^n$, $\nabla f(x) \in \mathbb{R}^n$

$\nabla^2 f(x) \in \mathbb{R}^{n \times n}$

↑ Hessian matrix (second-order derivatives)

↳ Because f is C^2, the second-order optimality conditions hold:

Second-order necessary conditions:

$$\left[\bar{x} \in \mathbb{R}^n \text{ is a local minimum of (P)}\right] \Rightarrow \begin{cases} \|\nabla f(\bar{x})\| = 0 \\ \lambda_{\min}\left(\nabla^2 f(\bar{x})\right) \geq 0 \end{cases}$$

$\Leftrightarrow \nabla^2 f(\bar{x}) \succeq 0$

$\Leftrightarrow \forall v \in \mathbb{R}^n$, $v^T \nabla^2 f(\bar{x}) v \geq 0$

NB: If f is convex, then $\lambda_{\min}\left(\nabla^2 f(x)\right) \geq 0 \quad \forall x \in \mathbb{R}^n$

Second-order sufficient conditions

$$\left.\begin{array}{l} \|\nabla f(\bar{x})\| = 0 \\ \lambda_{\min}\left(\nabla^2 f(\bar{x})\right) > 0 \end{array}\right\} \Rightarrow \left[\bar{x} \text{ is a local minimum of (P)}\right]$$

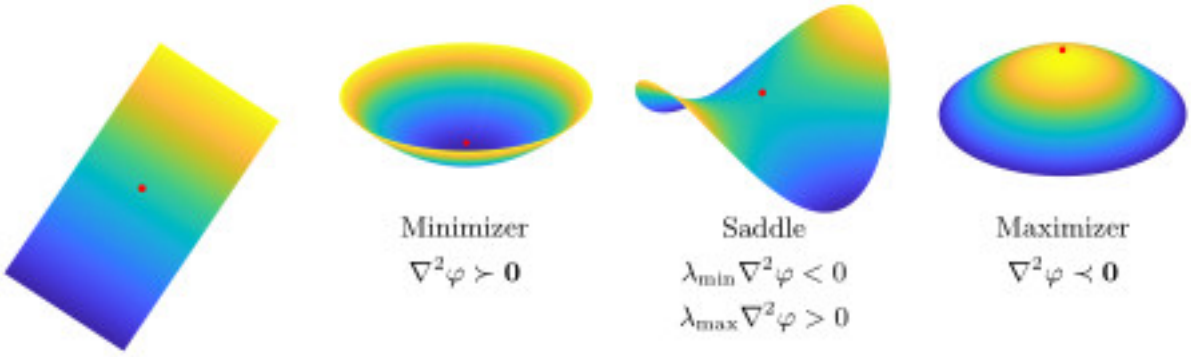
↓ Always true for strongly convex f

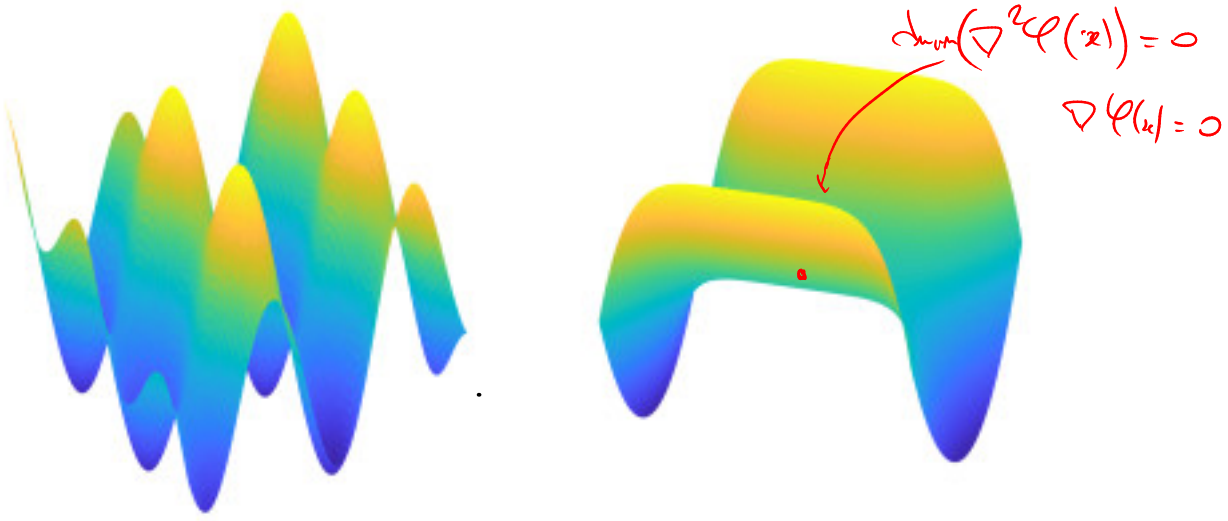
→ In the nonconvex setting, second-order information is

helpful to classify first-order stationary points ($\bar{x}$ such that $\|\nabla f(\bar{x})\| = 0$)

aka critical points

Can have many different kinds of critical points





Local minimum ≠ global minimum

Takeaway: Converging to a point $\bar{x} \in \mathbb{R}^n$ such that

$\|\nabla f(\bar{x})\| = 0$ and $\lambda_{\min}(\nabla^2 f(\bar{x})) \geq 0$

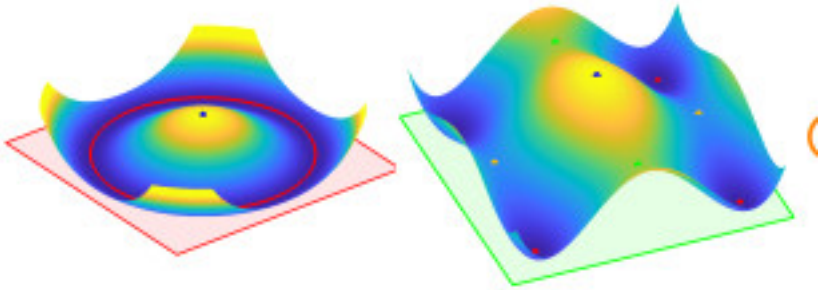
does not mean that you converge to a global minimum (not even a local one)

BUT: → the second-order conditions can be checked in polynomial time and they are the strongest guarantee that one can define for C^2 problems

→ In certain classes of nonconvex problems,

$$\left[\begin{array}{l}\|\nabla f(\bar{x})\| = 0 \\ \nabla^2 f(\bar{x}) \succeq 0\end{array}\right] \Leftrightarrow \left[\bar{x} \in \operatorname*{argmin}_x f(x)\right]$$

i.e. the 2nd order necessary conditions characterize global optima!



Objectives for low-rank matrix completion problems

Example: rank-1 matrix completion

minimize $u \in \mathbb{R}^n$ $\quad \frac{1}{2}\|uu^T - M\|_F^2 = \sum_i \sum_j \left([uu^T]_{ij} - M_{ij}\right)^2$

(Plots from J. Wright & Y. Ma (2022) "High-dimensional data analysis with low-dimensional models")

→ what about complexity ?

Def: $\bar{x} \in \mathbb{R}^n$ is an $(\varepsilon_g, \varepsilon_H)$-point for (P) if

$$\|\nabla f(\bar{x})\| \leq \varepsilon_g \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\bar{x})) \geq -\varepsilon_H$$

where $\varepsilon_g, \varepsilon_H$ are positive tolerances (typically in $(0,1)$)

Goal: Design algorithms that provably reach an $(\varepsilon_g, \varepsilon_H)$-point at a cost polynomial in ε_g^{-1} and ε_H^{-1}

(2) Basic method: Gradient descent + Negative curvature

Recall: Gradient descent $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ $\forall k \geq 0$

$\|\nabla f(x_k)\| \leq \varepsilon_g$ after $O(\varepsilon_g^{-2})$ iterations

Algorithm ($x_0 \in \mathbb{R}^n$, $k=0$, $\varepsilon_g, \varepsilon_H$)

↑ optional (can define the method without those)

(1) Compute $\nabla f(x_k)$.

(2) If $\|\nabla f(x_k)\| \geq \varepsilon_g$

Set $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$. for some $\alpha_k > 0$

(3) Else if $\lambda_{\min}(\nabla^2 f(x_k)) < -\varepsilon_H$

Compute $v_k \in \mathbb{R}^n$ such that ($v_k \neq 0$!)

$$\begin{cases} \nabla f(x_k)^T v_k \leq 0 \\ \nabla^2 f(x_k) v_k = \lambda_{\min}(\nabla^2 f(x_k)) v_k \end{cases}$$

and then set $x_{k+1} = x_k + \alpha_k v_k$ for some $\alpha_k > 0$

(4) Otherwise ($\|\nabla f(x_k)\| \leq \varepsilon_g$ and $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\varepsilon_H$) STOP.

(5) Set $k = k+1$ and go back to (1)

Analysis

- Assumptions: 1) $f \in C^{1,1}_{L_g}$

$$\Rightarrow f(x+s) \leq f(x) + \nabla f(x)^T s + \frac{L_g}{2}\|s\|^2$$

$\forall (x,s)$ in $(\mathbb{R}^n)^2$

2) $f \in C^{2,2}_{L_H}$

$$\Rightarrow f(x+s) \leq f(x) + \nabla f(x)^T s + \frac{1}{2} s^T \nabla^2 f(x) s + \frac{L_H}{6}\|s\|^3$$

$\forall (x,s) \in (\mathbb{R}^n)^2$

3) $f(x) \geq f_{low} \in \mathbb{R}$ $\forall x \in \mathbb{R}^n$

↳ Key lemma

If $\|\nabla f(x_k)\| \geq \varepsilon_g$, then with $\alpha_k = \frac{1}{L_g}$,

$$f(\underbrace{x_k - \alpha_k \nabla f(x_k)}_{x_{k+1}}) \leq f(x_k) - \frac{1}{2L_g}\|\nabla f(x_k)\|^2$$

$$\leq f(x_k) - \frac{1}{2L_g}\varepsilon_g^2$$

$\Rightarrow$ The method will find x_k such that $\|\nabla f(x_k)\| \leq \varepsilon_g$ in at most $\mathcal{O}(\varepsilon_g^{-2})$ iterations

- If $\lambda_{\min}(\nabla^2 f(x_k)) < -\varepsilon_H$ (and $\|\nabla f(x_k)\| \leq \varepsilon_g$)

$$f(x_{k+1}) = f(x_k + \alpha_k v_k)$$

$$f \overset{C^{2,2}}{\underset{L_H}{\longrightarrow}} \leq f(x_k) + \alpha_k \underbrace{\nabla f(x_k)^T v_k}_{\leq 0} + \frac{\alpha_k^2}{2} \underbrace{v_k^T \nabla^2 f(x_k) v_k}_{\uparrow} + \frac{L_H}{6}\alpha_k^3 \|v_k\|^3$$

$$v_k^T \lambda_{\min}(\nabla^2 f(x_k)) v_k = \lambda_{\min}(\nabla^2 f(x_k)) \|v_k\|^2$$

$$= f(x_k) + \frac{\lambda_{\min}(\nabla^2 f(x_k)) \|v_k\|^2 \alpha_k^2}{2} + \frac{L_H}{6}\alpha_k^3 \|v_k\|^3$$

$$\leq f(x_k) - \alpha_k^2 \frac{\|v_k\|^2}{2} \varepsilon_H + \frac{L_H}{6}\alpha_k^3 \|v_k\|^3$$

Setting $\alpha_k = \dfrac{\varepsilon_H}{L_H \|v_k\|} > 0$ gives

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L_H^2}\varepsilon_H^3 + \frac{1}{6L_H^2}\varepsilon_H^3$$

$$= f(x_k) - \frac{1}{3L_H^2}\varepsilon_H^3$$

$\Rightarrow$ The method finds x_k such that $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\varepsilon_H$ after at most $\mathcal{O}(\varepsilon_H^{-3})$ iterations

↳ To get a complexity bound, observe that

x_k not $(\varepsilon_g, \varepsilon_H)$-point $\Rightarrow$ $f(x_{k+1}) \leq f(x_k) - \min\left(\frac{\varepsilon_g^2}{2L_g}, \frac{\varepsilon_H^3}{3L_H^2}\right)$

As a result, the method finds an $(\varepsilon_g, \varepsilon_H)$-point in at most $O\left(\max\left(\varepsilon_g^{-2}, \varepsilon_H^{-3}\right)\right)$ iterations

Optimal complexity for $C^{1,1}_{L_g}$ and $C^{2,2}_{L_H}$ functions:

$$O\left(\max\left(\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\right)\right)$$

→ Achieved by "convex until proven guilty" method from lecture 2

→ Also achieved by certain Newton-type methods

(3) Second-order algorithms: Trust-region methods and cubic regularization

Starting point: $f \in C^2$, so we want to use Newton's method.

$$x_{k+1} = x_k + s_k, \qquad \nabla^2 f(x_k)\, s_k = -\nabla f(x_k)$$

Issue: The Hessian matrix is not necessarily invertible or positive definite, so the method is not well-defined (let alone equipped with complexity bounds) for non-convex f

↳ Newton-type methods (aka second-order methods) globalize Newton's method to guarantee well-definiteness and convergence

Ex) Line search, Regularization, Trust region

≈ GD+Negative curvature — Adaptive methods

Trust-region method

Inputs: $x_0 \in \mathbb{R}^n$, $\delta_0 > 0$, $\delta_{max} \geq \delta_0$, $\eta > 0$

For $k = 0, 1, 2, \ldots$

- Compute $s_k \in \mathbb{R}^n$ as a solution of

Trust-region subproblem:

$$\begin{cases} \text{minimize}_{s \in \mathbb{R}^n} & f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s \\ \text{subject to} & \|s\| \leq \delta_k \end{cases}$$

Trust-region model of f around x_k (= Taylor expansion)

Trust region: $\{s \mid \|s\| \leq \delta_k\}$

- If $\left(f(x_k) - f(x_k + s_k)\right) \geq \eta \left(f(x_k) - \left(f(x_k) - \nabla f(x_k)^T s_k - \frac{1}{2} s_k^T \nabla^2 f(x_k) s_k\right)\right)$

  actual change in f ; model at $s = 0$; model at $s = s_k$

  Set $x_{k+1} = x_k + s_k$ and $\delta_{k+1} = \min(2\delta_k, \delta_{max})$

Else

  Set $x_{k+1} = x_k$ and $\delta_{k+1} = \delta_k / 2$

Key features:

- Acceptance condition: guarantees $f(x_{k+1}) < f(x_k)$ when $x_{k+1} \neq x_k$
- Adaptive parameter (δ_k: trust-region radius)

$\{\delta_k\}$ drives the convergence process

- Subproblem always has a solution (even when Newton's method is not well defined)
  - → If $\nabla^2 f(x_k)$ is invertible and δ_k is sufficiently large, s_k is the Newton step!
  - → Possible to find an approximate solution in polynomial time, can be solved explicitly in small dimension

---

Complexity analysis (Goal: Find x_k such that $\|\nabla f(x_k)\| \leq \varepsilon_g$ and $\lambda_{\min}(\nabla^2 f(x_k)) \geq -\varepsilon_H$)

→ If $\|\nabla f(x_k)\| > \varepsilon_g$ and $x_{k+1} = x_k + s_k$, ($\|s_k\| = \delta_k$)

$$f(x_k) - f(x_{k+1}) \geq O\left(\min\left(\delta_k^2, \|\nabla f(x_k)\| \delta_k\right)\right)$$

($\|s_k\| \leq \delta_k$)

$$\geq O\left(\min(\delta_k^2, \varepsilon_g \delta_k)\right)$$

Argument: s_k is a solution of the subproblem, so it decreases the model at least as much as any step of the form $\alpha \nabla f(x_k)$ with $|\alpha| \|\nabla f(x_k)\| \leq \delta_k$

→ If $\lambda_{\min}(\nabla^2 f(x_k)) < -\varepsilon_H$ and $x_{k+1} = x_k + s_k$, ($\|s_k\|^2 = \delta_k^2$)

$$f(x_k) - f(x_{k+1}) \geq O\left(|\lambda_{\min}(\nabla^2 f(x_k))| \delta_k^2\right)$$

$$\geq O(\varepsilon_H \delta_k^2)$$

Reasoning: s_k decreases the model as much as a step in a negative curvature direction (always a boundary step of norm δ_k)

→ Overall, for every successful iteration ($x_{k+1} = x_k + s_k$), if x_k is not an $(\varepsilon_g, \varepsilon_H)$-point, we have

$$f(x_k) - f(x_{k+1}) \geq O\left(\min\left(\delta_k^2, \varepsilon_g \delta_k, \varepsilon_H \delta_k^2\right)\right)$$

→ Meanwhile, using properties of f and the rules for updating δ_k, can show:

- $\delta_k \geq O(\min(\varepsilon_g, \varepsilon_H))$ $\forall k$ as long as an $(\varepsilon_g, \varepsilon_H)$-point has not been found.

$S_k = \{j < k \mid x_{j+1} \neq x_j\}$
$U_k = \{j < k \mid x_{j+1} = x_j\}$

$|U_k| \leq C |S_k|$

- The number of unsuccessful iterations ($x_{k+1} = x_k$) is bounded above by a constant times the number of successful iterations

⇒ For trust region and other adaptive methods, it suffices to bound the number of successful iterations to get a complexity bound.

⇒ To bound the number of successful iterations, we use

$$f(x_0) - f_{low} \geq f(x_0) - f(x_k) = \sum_{j=0}^{k-1} \left(f(x_j) - f(x_{j+1})\right) = \sum_{j \in S_k} \left(f(x_j) - f(x_{j+1})\right)$$

$$\geq \sum_{j \in S_k} O\left(\min\left(\delta_k^2, \varepsilon_g \delta_k, \varepsilon_H \delta_k^2\right)\right)$$

$$\geq \sum_{j \in S_k} O\left(\min\left(\varepsilon_g^2, \varepsilon_H^2, \varepsilon_g \varepsilon_H, \varepsilon_H \varepsilon_g^2, \varepsilon_H^3\right)\right)$$

$\varepsilon_g, \varepsilon_H \in (0,1) \longrightarrow \left(\geq\right) \sum_{j \in S_k} O\left(\min\left(\varepsilon_g^2 \varepsilon_H, \varepsilon_H^3\right)\right)$

$\Rightarrow$ Complexity: the trust-region method reaches an $(\varepsilon_g, \varepsilon_H)$-point in at most $O\left(\max\left(\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\right)\right)$ iterations

GD + Negative curvature: $O\left(\max\left(\varepsilon_g^{-2}, \varepsilon_H^{-3}\right)\right)$

Trust region (2012): $O\left(\max\left(\varepsilon_g^{-2}\varepsilon_H^{-1}, \varepsilon_H^{-3}\right)\right)$ → worse than the first bound

→ sharp (∃ f on which trust region attains the lower bound)

$\Rightarrow$ these results are really contradictory with the practice, where Trust Region works much better than GD technique (same observation than GD vs Newton)

Cubic regularization

- Introduced in 1981 by Griewank (unpublished technical report, no complexity)
- Revival: Nesterov & Polyak 2006
  - $\Rightarrow$ Got the $O\left(\max\left(\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\right)\right)$ complexity
- Cartis Gould Toint 2011
  - $\Rightarrow$ Adaptive version of the 2006 method

Algorithm ($x_0 \in \mathbb{R}^n$, $\eta > 0$, $0 < \sigma_{min} \leq \sigma_0$)

for $k = 0, 1, \ldots$

- Compute s_k as a solution of

cubic subproblem [$\underset{s \in \mathbb{R}^n}{\text{minimize}} \underbrace{\{ f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s}_{m_k(s)} + \underbrace{\frac{\sigma_k}{3} \|s\|^3}_{\text{cubic regularization term}} \}$

- If $f(x_k) - f(x_k + s_k) \geq \eta (m_k(0) - m_k(s))$

$x_{k+1} = x_k + s_k$ and $\sigma_{k+1} = \max\left(\frac{\sigma_k}{2}, \sigma_{min} \right)$

Otherwise

$x_{k+1} = x_k$ and $\sigma_{k+1} = 2\sigma_k$.

Key differences with trust region

1) subproblem: both involve $q_k(s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s$

→ TR: minimize under $\|s\| \leq \delta_k$ (explicit regularization)

→ CR (cubic): ——— a regularized version of q_k

2) σ_k is $\approx$ equivalent to $\frac{1}{\delta_k}$ (actually $\frac{1}{\delta_k \|s_k\|}$)

Complexity: Analysis very similar to that of trust-region

- $\sigma_k \leq O(1)$ $\forall k$ as long as an $(\varepsilon_g, \varepsilon_H)$-point has not been found
- # of unsuccessful iterations $\leq O(1) \times$ # successful iterations

- If successful iteration,

$$f(x_k) - f(x_{k+1}) \geq O\left(\|s_k\|^3\right)$$

and if x_{k+1} is not an $(\varepsilon_g, \varepsilon_H)$-point,

then

$$\|s_k\| \geq O\left(\min\left(\|\nabla f(x_{k+1})\|^{1/2}, -\lambda_{\min}(\nabla^2 f(x_{k+1}))\right)\right)$$

$$\left(\geq O\left(\min\left(\varepsilon_g^{1/2}, \varepsilon_H\right)\right)\right.$$

Really distinctive part

$$\|s_k\| \geq O\left(\|\nabla f(x_{k+1})\|^{1/2}\right)$$

$\Rightarrow$ that result is typical of Newton-type methods

Overall,

$$f(x_0) - f_{low} \geq \sum_{j \in S_k} O\left(\|s_k\|^3\right)$$

$$\geq \sum_{j \in S_k} O\left(\min\left(\|\nabla f(x_{k+1})\|, -\lambda_{\min}(\nabla^2 f(x_{k+1}))\right|^3\right)$$

$$\geq \sum_{j \in S_k} O\left(\min\left(\varepsilon_g^{-3/2}, \varepsilon_H^3\right)\right)$$

$\Rightarrow$ Complexity $O\left(\max\left(\varepsilon_g^{-3/2}, \varepsilon_H^{-3}\right)\right)$: optimal among 2nd-order methods

Trust Region Vs Cubic regularization

Bad complexity
Very good implementations

Optimal complexity
Very few efficient implementations

→ Several attempts to:

* Make cubic regularization practical
  (still challenging because the subproblem is not that easy to solve)
* Improve the complexity of TR without hurting practical performance
  - TRACE (Curtis, Robinson, Samadi, 2017)
  - TR-Newton (Curtis et al 2021)

Change to the original method

Compute s_k as the solution to

$$\underset{s \in \mathbb{R}^n}{\text{minimize}} \; f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \left(\nabla^2 f(x_k) + \varepsilon_H I \right) s$$

$$\text{s.t.} \quad \|s\| \leq \delta_k$$