Modern analysis of local convergence for classical quasi-Newton methods

Anton Rodomanov

UCLouvain, Belgium

March 13, 2023

Toulouse School of Economics, Toulouse

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Optimization Problems

Minimize a given function subject to certain constraints:

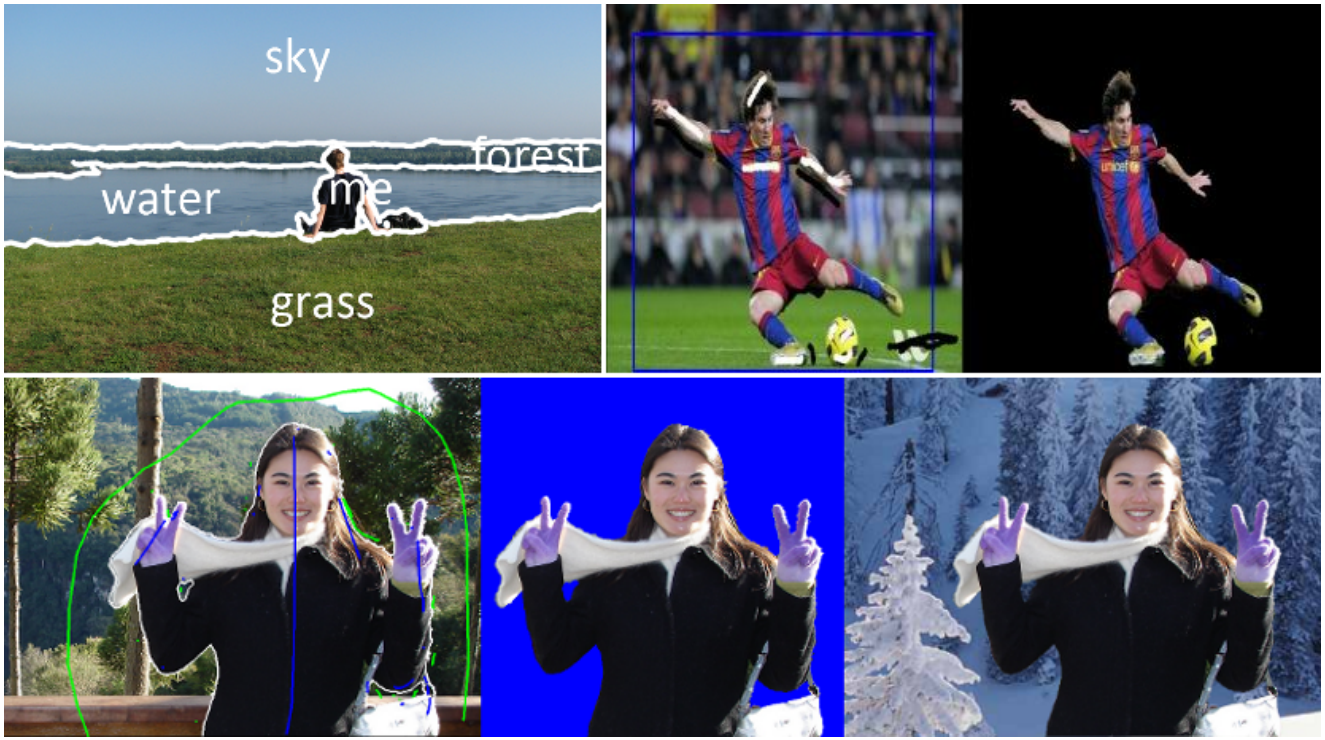
$$\min_{x \in Q} f(x),$$

where $Q \subseteq \mathbb{R}^n$.

Many applications:

- Machine Learning
- Economics
- Engineering
- Telecommunications
- Signal Processing
- ...

Example: Machine Learning



Empirical loss minimization:

$$\min_x \sum_{i=1}^{m} \ell\big(b_i, \hat{b}(x, a_i)\big)$$

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Gradient Method

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

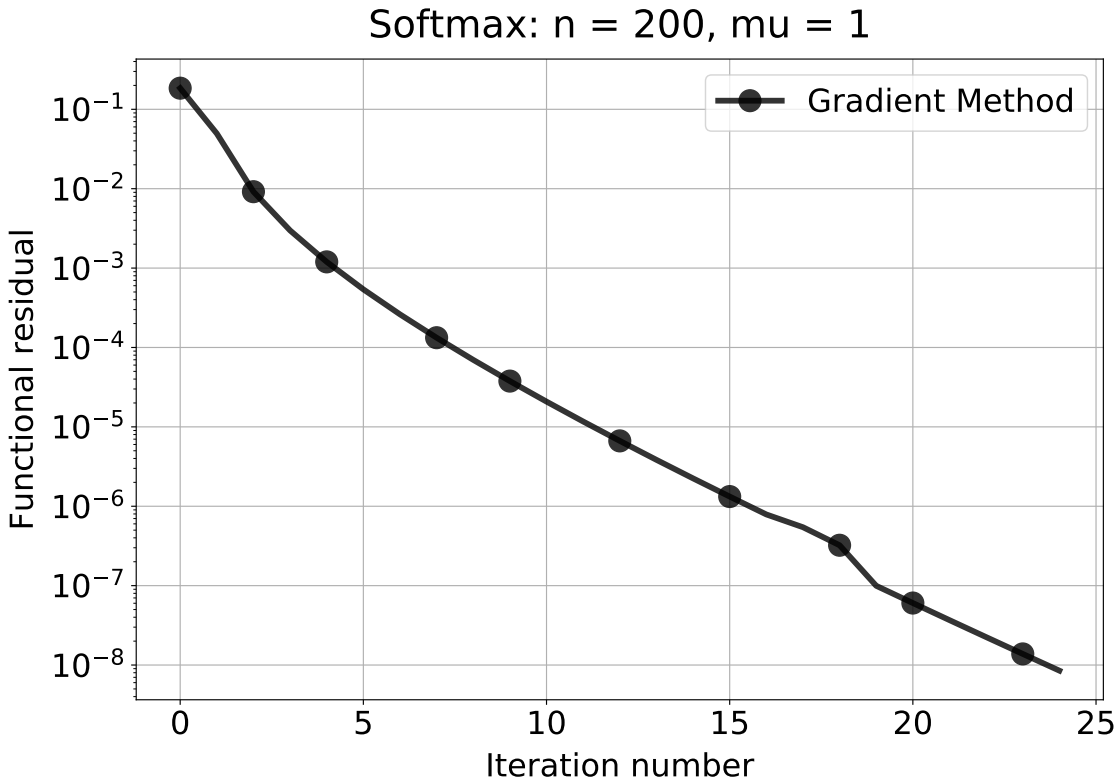
Gradient of a function: $\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_i}\right)_{i=1}^{n} \ (\in \mathbb{R}^n)$.

Gradient Method

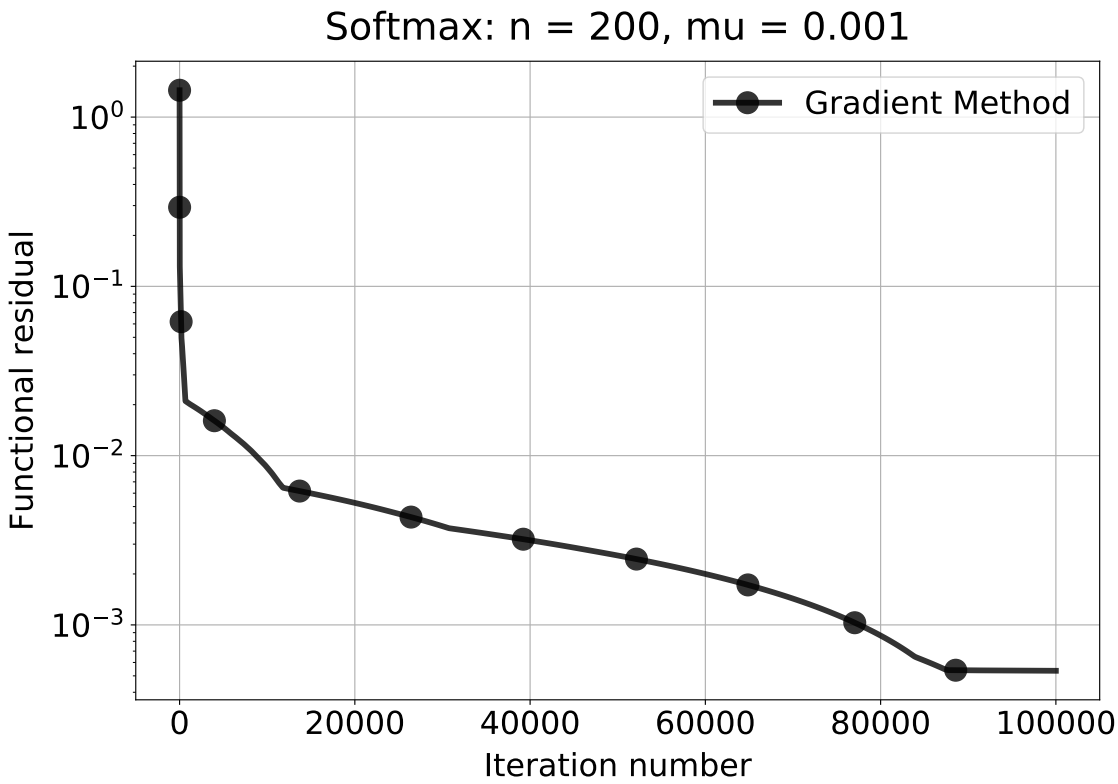
$$x_{k+1} = x_k - \alpha_k \nabla f(x_k), \qquad k \geq 0,$$

where $\alpha_k \geq 0$ are certain "step sizes".

Example: Convergence Plot



Example: Slow Convergence



Why so slow?

Convergence Theory for Gradient Method

Problem class: Strongly convex functions with Lipschitz gradient:

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \forall x \in \mathbb{R}^n,$$

where $\nabla^2 f(x) = \left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)_{i,j=1}^n \quad (\in \mathbb{S}^n)$.

Main parameter: Condition number $\varkappa := \frac{L}{\mu}$ (≥ 1).

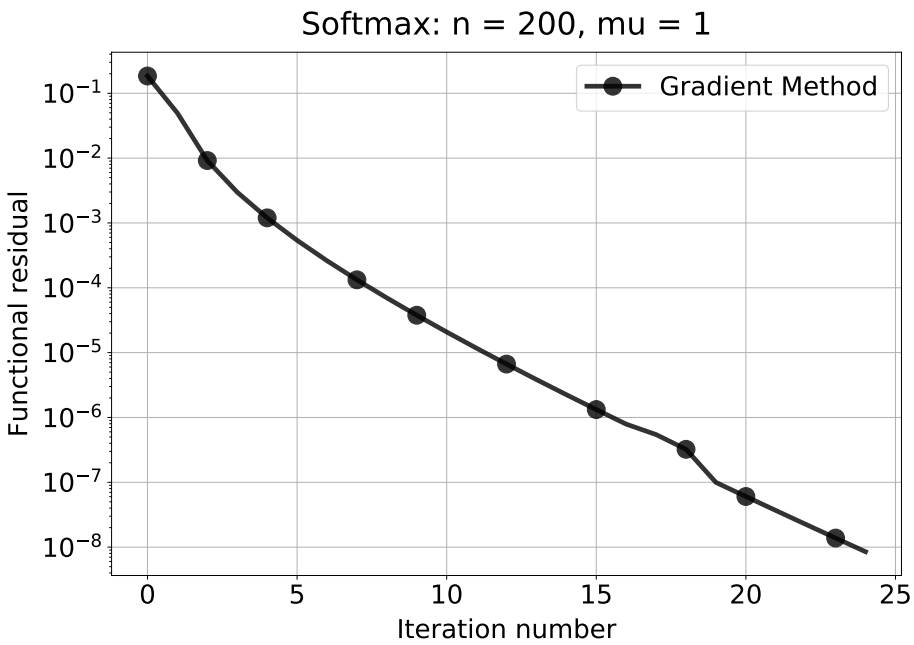
Convergence rate: $f(x_k) - f^* \leq (1 - \varkappa^{-1})^k [f(x_0) - f^*]$.

Complexity bound

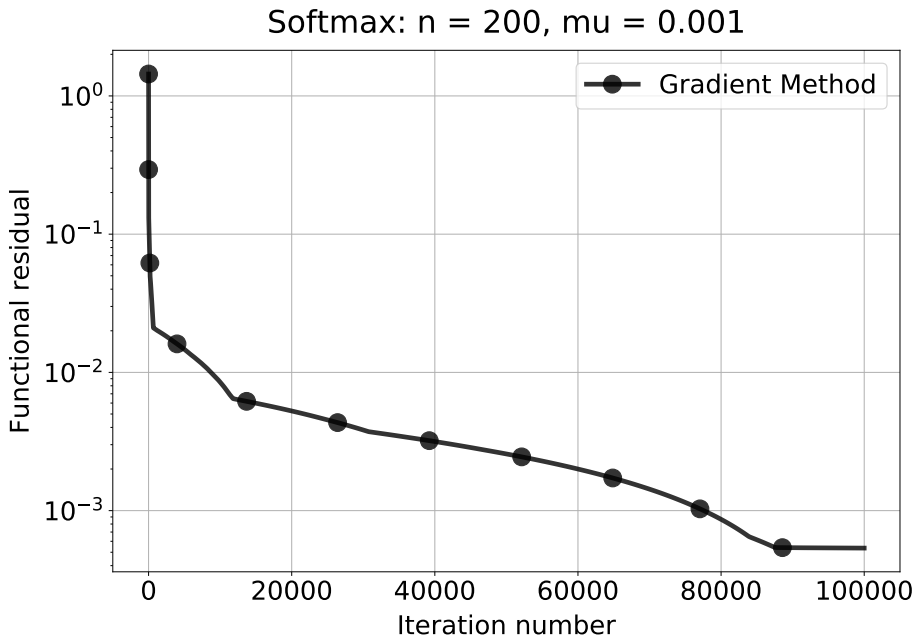
$$K(\epsilon) = \varkappa \ln(\epsilon^{-1})$$

iterations to find x_k such that $f(x_k) - f^* \leq \epsilon [f(x_0) - f^*]$.

Good and Bad Examples Revisited



$\varkappa \approx 10$



$\varkappa \approx 10,000,000$

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Newton's Method

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

Newton's Method

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k).$$

Interpretation: Automatic preconditioning (scaling) of Gradient Method.

Main idea: Minimize quadratic model around previous point:

$$f(x) \approx f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle.$$

Newton's Method: Convergence Rate

Main assumption: f is μ-strongly convex with L_2-Lipschitz Hessian.

Local quadratic convergence:

$$\|\nabla f(x_{k+1})\| \leq \frac{L_2}{2\mu^2}\|\nabla f(x_k)\|^2.$$

Very fast convergence: $r_{k+1} \leq r_k^2$ $(0.1 \to 0.01 \to 0.0001 \to \dots)$.

Complexity bound

When started sufficiently close to solution[1], Newton's method needs

$$K(\epsilon) = \log_2 \log_2 O(\epsilon^{-1})$$

iterations to find x_k such that $f(x_k) - f^* \leq \epsilon[f(x_0) - f^*]$.

[1] Specifically, when $\|\nabla f(x_0)\| \leq O(\mu^2/L_2)$.

Globally Convergent Variants of Newton's Method

Damped Newton's Method:

$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k),$$

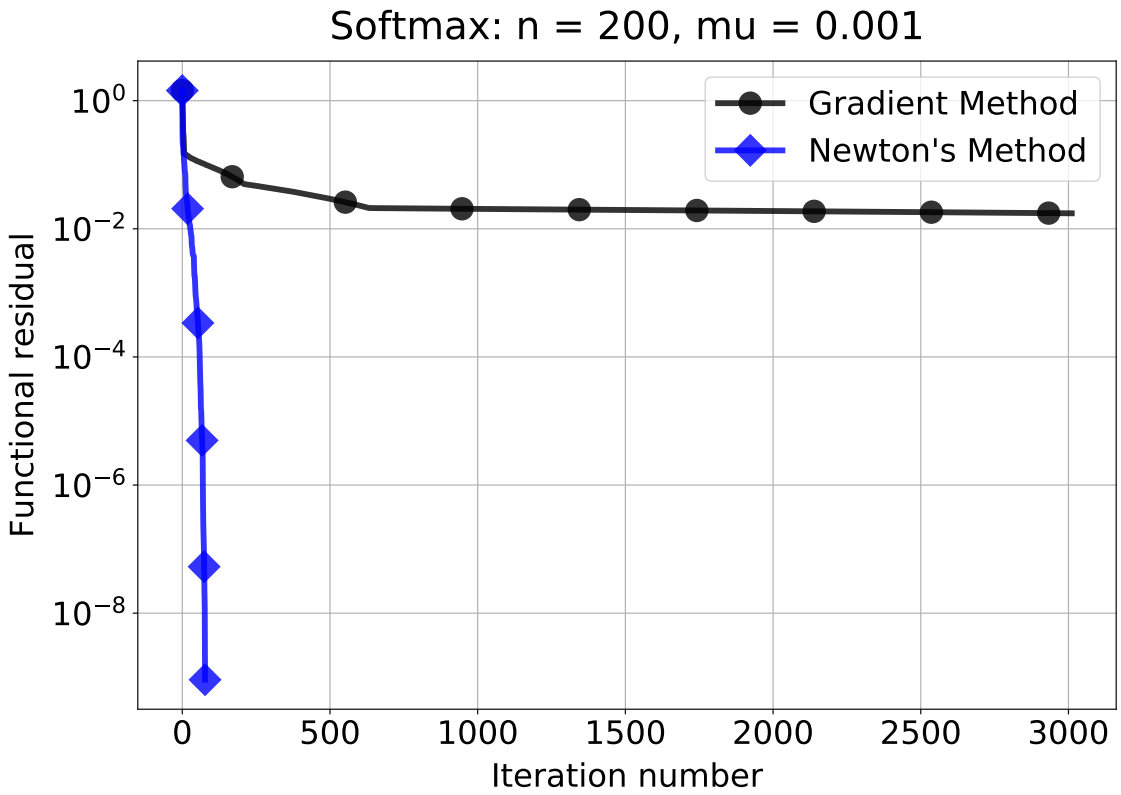
where $\alpha_k \in [0, 1]$ and $\alpha_k \to 1$ as $k \to \infty$.

Other variants:

- Levenberg-Marquardt regularization (Levenberg, 1944; Marquardt, 1963).
- Trust-region methods (Goldfeld et al., 1966; Conn et al., 2000).
- Cubic regularization (Nesterov and Polyak, 2006):

$$\begin{aligned} x_{k+1} &= \operatorname*{argmin}_{x \in \mathbb{R}^n} \Big\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ &\qquad + \frac{1}{2} \langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{M_k}{6} \|x - x_k\|^3 \Big\} \\ &= x_k - (\nabla^2 f(x_k) + \tfrac{1}{2} M_k r_k I)^{-1} \nabla f(x_k), \qquad r_k := \|x_{k+1} - x_k\|. \end{aligned}$$

Example


Softmax: n = 200, mu = 0.001
Gradient Method
Newton's Method
Functional residual
10^{0}
10^{-2}
10^{-4}
10^{-6}
10^{-8}
0
500
1000
1500
2000
2500
3000
Iteration number

Cost of One Iteration

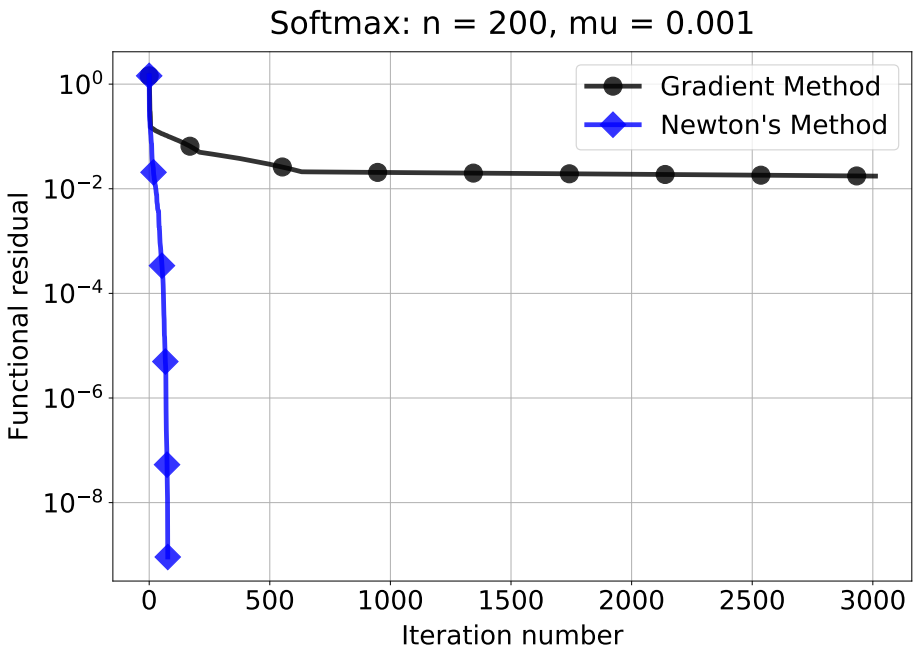
- **Gradient Method:** $x_{k+1} = x_k - \alpha_k \nabla f(x_k)$:

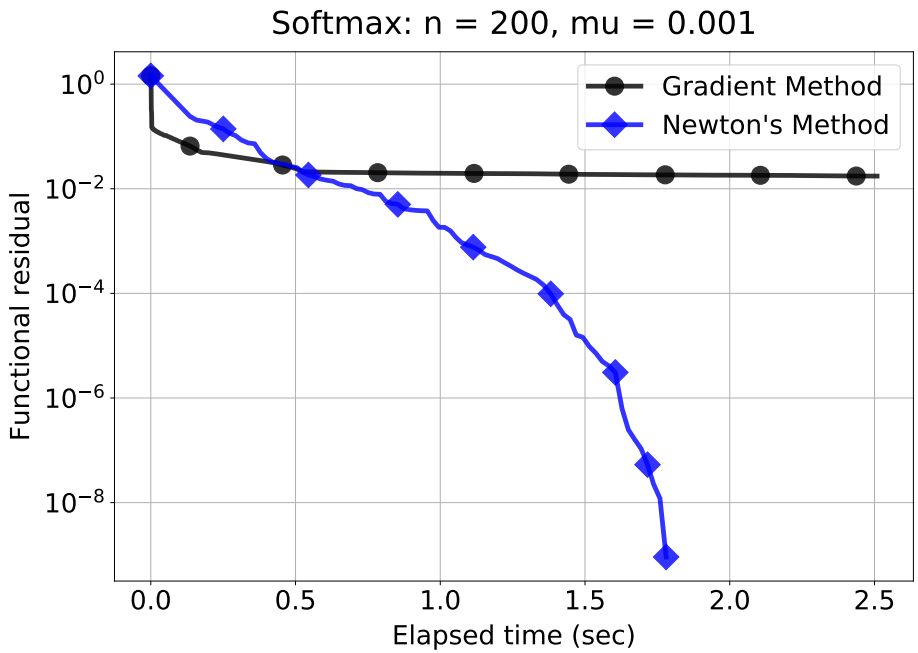
$$\text{Cost}(\nabla f) + O(n).$$

- **Newton's Method:** $x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$:

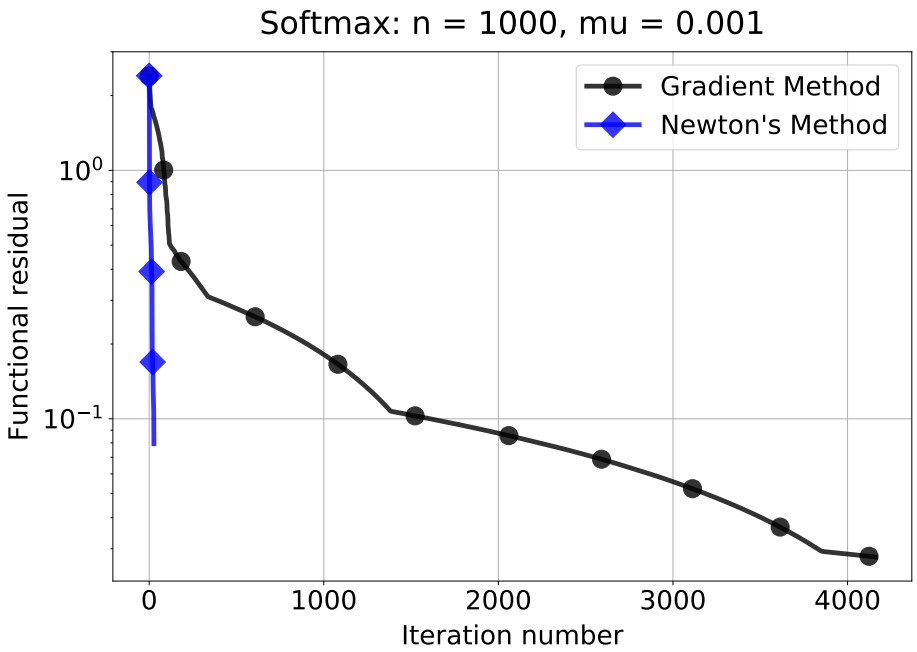
$$\text{Cost}(\nabla f) + \text{Cost}(\nabla^2 f) + O(n^3).$$

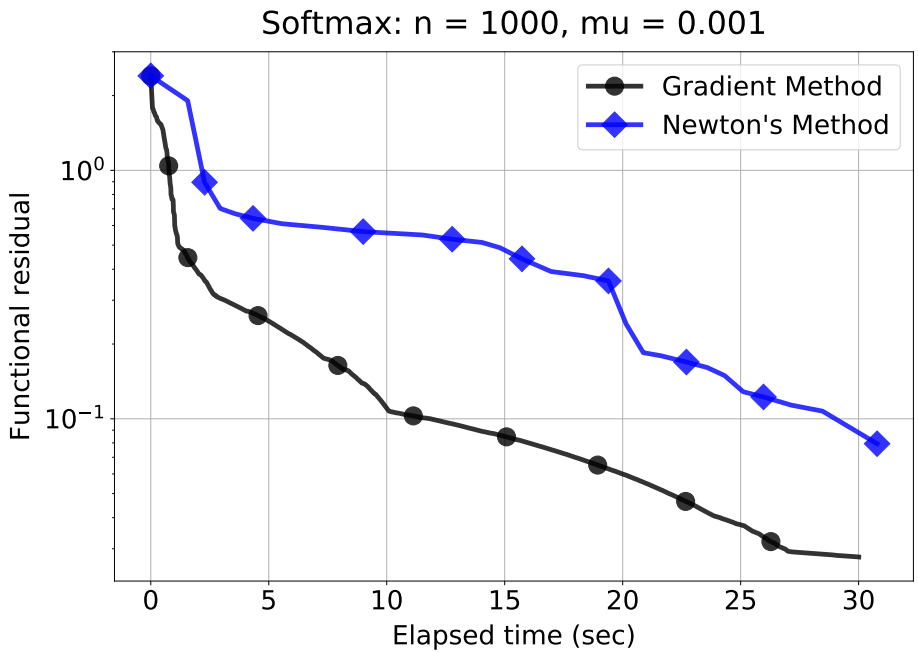
Example: Small Dimension





Example: Large dimension





Summary: Gradient Method vs Newton's Method

Method	Cheap iteration?	Fast convergence?
Gradient Method	Yes	No
Newton's Method	No	Yes

Can we have something in between?

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Quasi-Newton Methods (Davidon, 1959; Fletcher and Powell, 1963)

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

Quasi-Newton iteration

$$x_{k+1} = x_k - \alpha_k H_k \nabla f(x_k),$$

where

$$H_k \approx [\nabla^2 f(x_k)]^{-1}.$$

Main question: How to update H_k at each iteration?

Secant Equation

Goal: Update

$$H \approx [\nabla^2 f(x)]^{-1} \qquad \text{into} \qquad H_+ \approx [\nabla^2 f(x_+)]^{-1}$$

using the information computed at x and x_+:

$$\nabla f(x) \qquad \text{and} \qquad \nabla f(x_+).$$

Secant equation

Choose H_+ such that

$$H_+ \gamma = \delta, \tag{*}$$

where

$$\delta := x_+ - x, \qquad \gamma := \nabla f(x_+) - \nabla f(x).$$

Note: (*) is satisfied by J^{-1} for

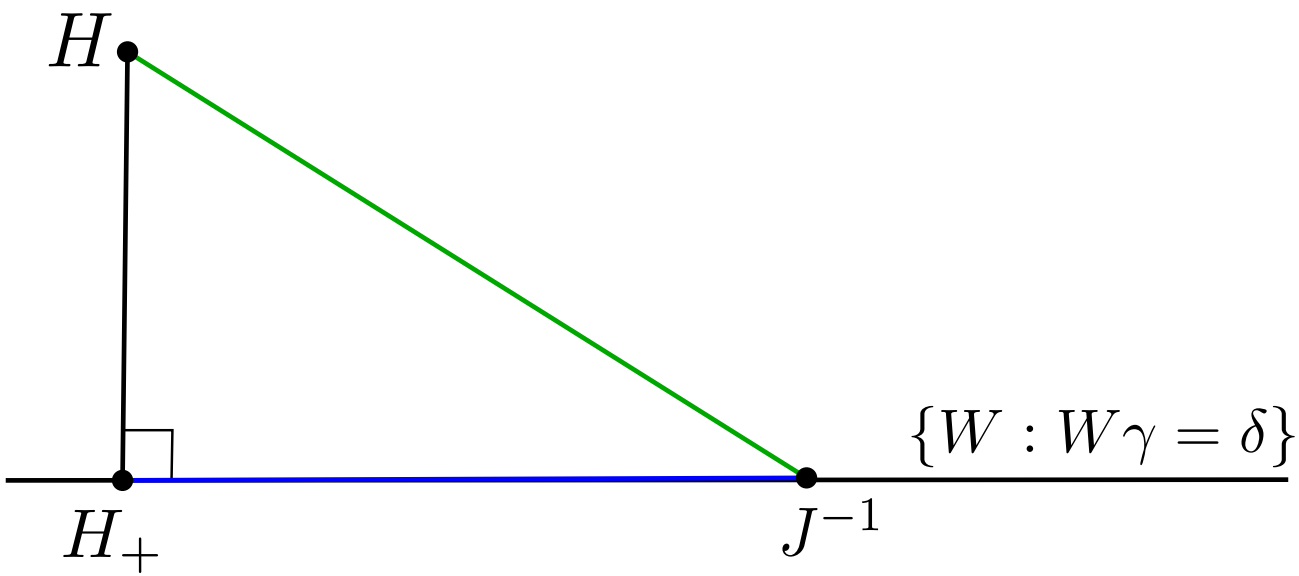
$$J := \int_0^1 \nabla^2 f(x + t\delta) dt \quad \left(\approx \nabla^2 f(x_+)\right).$$

Least Change Principle

Define some distance $\beta(\cdot, \cdot)$ between two (positive definite) matrices.

Least Change Problem

$$\min_{H_+} \{\beta(H, H_+) : H_+ \gamma = \delta\}.$$



Main Updating Formulas

- Davidon–Fletcher–Powell (DFP):

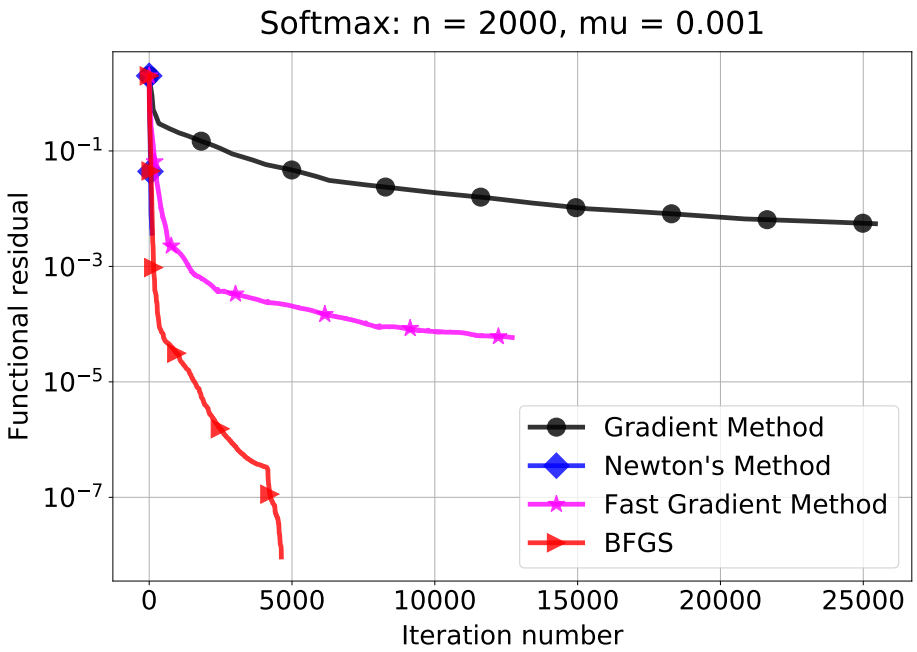
$$H_+ = \mathrm{DFP}^{-1}(H, \delta, \gamma) := H - \frac{H\gamma\gamma^T H}{\langle \gamma, H\gamma \rangle} + \frac{\delta\delta^T}{\langle \gamma, \delta \rangle}.$$

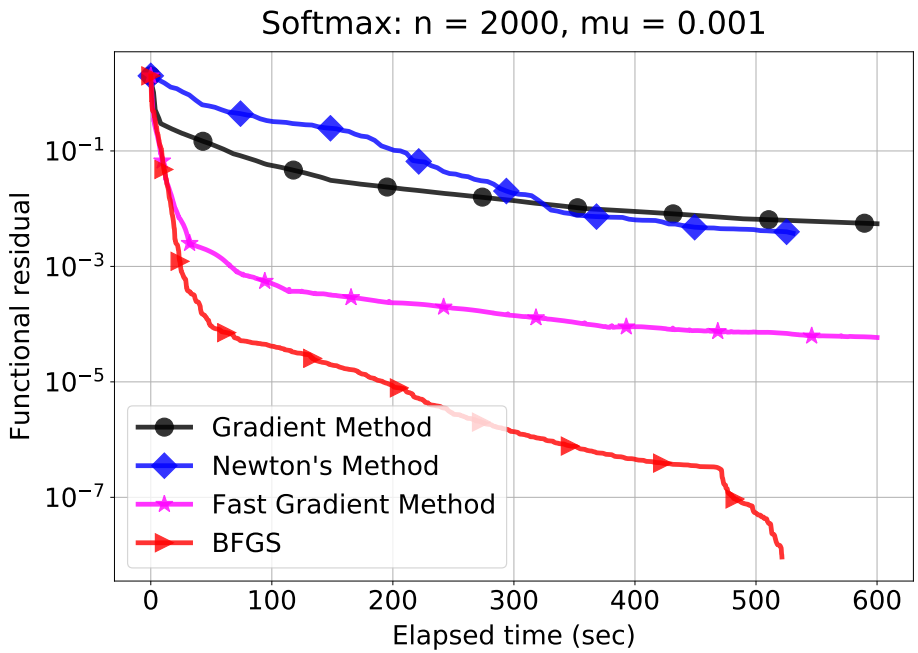
- Broyden–Fletcher–Goldfarb–Shanno (BFGS):

$$H_+ = \mathrm{BFGS}^{-1}(H, \delta, \gamma) := H - \frac{H\gamma\delta^T + \delta\gamma^T H}{\langle \gamma, \delta \rangle} + \left(\frac{\langle \gamma, H\gamma \rangle}{\langle \gamma, \delta \rangle} + 1 \right) \frac{\delta\delta^T}{\langle \gamma, \delta \rangle}.$$

Note: Cost of each update is $O(n^2)$ (not $O(n^3)$!).

Example


Softmax: n = 2000, mu = 0.001
Functional residual
10^{-1}
10^{-3}
10^{-5}
10^{-7}
0
5000
10000
15000
20000
25000
Iteration number
Gradient Method
Newton's Method
Fast Gradient Method
BFGS


Softmax: n = 2000, mu = 0.001
Functional residual
10^{-1}
10^{-3}
10^{-5}
10^{-7}
0
100
200
300
400
500
600
Elapsed time (sec)
Gradient Method
Newton's Method
Fast Gradient Method
BFGS

Summary: Quasi-Newton Methods

- Approximate Newton's Method without computing Hessian.
- Very efficient in practice.
- Can be extended to large-scale problems (L-BFGS).

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Classical Results on QN Methods

Local convergence (Broyden et al., 1973)

Suppose f is strongly convex with Lipschitz Hessian.
Consider either BFGS or DFP method with unit step sizes:

$$\alpha_k \equiv 1, \qquad \forall k \geq 0.$$

Then, $\forall \rho \in (0,1)$, $\exists \delta_1, \delta_2 > 0$ such that, $\forall (x_0, H_0)$ satisfying

$$\|x_0 - x^*\| \leq \delta_1 \qquad \text{and} \qquad \|H_0 - [\nabla^2 f(x^*)]^{-1}\| \leq \delta_2,$$

it holds that

$$\|x_{k+1} - x^*\| \leq \rho \|x_k - x^*\|, \qquad \forall k \geq 0.$$

Moreover, the rate of convergence is superlinear:

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

Classical Results on QN Methods II

Global convergence (Powell, 1976; Byrd and Nocedal, 1989)

Suppose f is strongly convex with Lipschitz gradient and Hessian.
Consider BFGS method with an appropriate line search[2].
Then, for any x_0 and H_0, it holds that

$$\lim_{k\to\infty} x_k = x^*.$$

Moreover, the rate of convergence is superlinear:

$$\lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0.$$

[2] One possible option is the standard backtracking line search: find the smallest integer $i_k \geq 0$ such that $\alpha_k = 2^{-i_k}$ satisfies

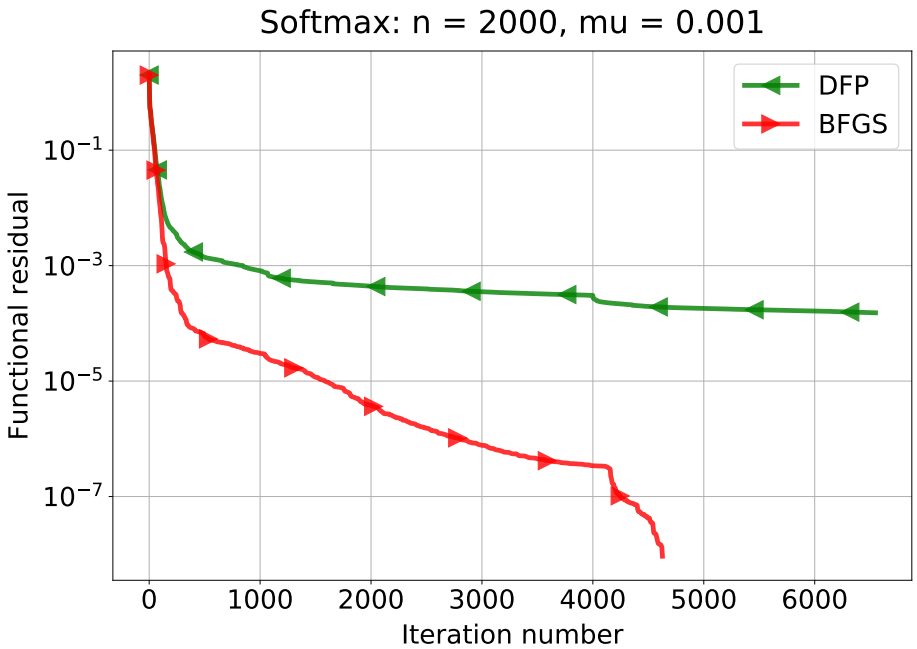
$$f(x_k - \alpha_k H_k \nabla f(x_k)) \leq f(x_k) - c_1 \alpha_k \langle \nabla f(x_k), H_k \nabla f(x_k) \rangle$$

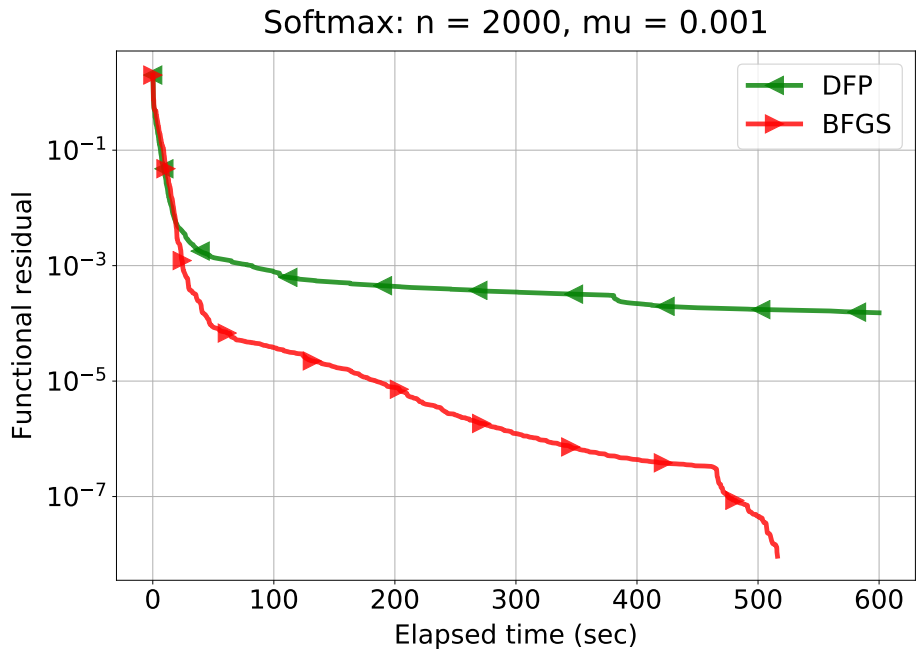
for a certain fixed constant $c_1 \in (0, 1)$.

Criticism

- Classical results are only qualitative (nonexplicit and asymptotic).
- No concrete efficiency estimates / complexity bounds.

Example (BFGS vs DFP):





- Classical results do not explain why BFGS is so much better.
- Cannot use them to compare BFGS with other methods.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Based on

Paper

A. Rodomanov and Y. Nesterov. New Results on Superlinear Convergence of Classical Quasi-Newton Methods. *Journal of Optimization Theory and Applications*, 188:744–769, 2021.

Related:

- A. Rodomanov and Y. Nesterov. Rates of superlinear convergence for classical quasi-Newton methods. *Mathematical Programming*, 194:159–190, 2022.
- A. Rodomanov and Y. Nesterov. Greedy Quasi-Newton Methods with Explicit Superlinear Convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Classical QN Methods

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

Convex Broyden Class ($\tau \in [0, 1]$)

$$\mathrm{Broyd}_\tau^{-1}(H, \delta, \gamma) := (1 - \tau)\,\mathrm{BFGS}^{-1}(H, \delta, \gamma) + \tau\,\mathrm{DFP}^{-1}(H, \delta, \gamma).$$

Remarks:

- Contains both BFGS ($\tau = 0$) and DFP ($\tau = 1$).
- Can be computed in $O(n^2)$ operations.

Convex Broyden Method

Choose $x_0 \in \mathbb{R}^n$, $H_0 \in \mathbb{S}^n_{++}$. Iterate for $k \geq 0$:

1. Set $x_{k+1} := x_k - H_k \nabla f(x_k)$.
2. Compute $\delta_k := x_{k+1} - x_k$ and $\gamma_k := \nabla f(x_{k+1}) - \nabla f(x_k)$.
3. Update $H_{k+1} := \mathrm{Broyd}_\tau^{-1}(H_k, \delta_k, \gamma_k)$.

Problem Class

1. f is μ-strongly convex with L-Lipschitz gradient ($\mu, L > 0$):

$$\mu I \preceq \nabla^2 f(x) \preceq LI, \qquad \forall x \in \mathbb{R}^n.$$

   Condition number: $\varkappa := \frac{L}{\mu}$ (≥ 1).

2. f is M-strongly self-concordant ($M \geq 0$):

$$\nabla^2 f(x) - \nabla^2 f(y) \preceq M\|x - y\|_z \nabla^2 f(w), \qquad \forall x, y, z, w \in \mathbb{R}^n,$$

   where $\|h\|_z := \langle \nabla^2 f(z)h, h\rangle^{1/2}$.

Remarks:

- Strong self-concordance $\implies$ self-concordance.
- 1 + 2 $\iff$ 1 + L_2-Lipschitz Hessian ($M = L_2/\mu^{3/2}$).
- 2 is an affine invariant property.
- For quadratic functions, $M = 0$.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Final Complexity Bound

Main quantity: "Starting moment of superlinear convergence"

$$K_0 = O\big(n\varkappa_\tau \ln(2\varkappa)\big), \qquad \varkappa_\tau := (1-\tau+\tau\tfrac{4}{9}\varkappa^{-1})^{-1}.$$

Assumptions:

- Initial point x_0 is suff. good: $M\|\nabla f(x_0)\|_{x_0}^* \le \max\{O(\varkappa^{-1}), K_0^{-1}\}$.
- Initial Hessian approximation: $H_0 := \frac{1}{L} I$.

Complexity bound

$$K(\epsilon) = \min\Big\{ \underbrace{\varkappa \ln O(\epsilon^{-1})}_{\substack{\text{Complexity of}\\ \text{Gradient method}}}, K_0 + \ln O(\epsilon^{-1})\Big\},$$

iterations to produce x_k such that $f(x_k) - f^* \le \epsilon[f(x_0) - f^*]$.

Discussion

Method	Complexity
BFGS	$\min\left\{\varkappa \ln O(\epsilon^{-1}), O\left(n \ln(2\varkappa)\right) + \ln O(\epsilon^{-1})\right\}$
DFP	$\min\left\{\varkappa \ln O(\epsilon^{-1}), O\left(n\varkappa \ln(2\varkappa)\right) + \ln O(\epsilon^{-1})\right\}$

- BFGS is almost insensitive to condition number $\varkappa$.
- For $\varkappa \gg n$, its total arithmetical complexity is essentially

$$O\left(n \ln(2\varkappa)\right) \times O(n^2) = O\left(n^3 \ln(2\varkappa)\right) = \widetilde{O}(n^3)$$

  (similar to one Newton's step).
- In contrast, for ill-conditioned problems ($\varkappa \gg n$), the superlinear convergence of DFP may be of no practical use.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Main Result

Local gradient norm: $\lambda_k := \|\nabla f(x_k)\|^*_{x_k}$.

Theorem. Let $H_0 = \frac{1}{L} I$ and x_0 be sufficiently close to solution:

$$M\lambda_0 \leq \frac{\ln(3/2)}{(3/2)^{3/2}} \max\{(2\varkappa)^{-1}, (K_0+9)^{-1}\}, \quad K_0 := \lceil 8n\varkappa_\tau \ln(2\varkappa)\rceil,$$

where $\varkappa_\tau := (1-\tau+\tau\frac{4}{9}\varkappa^{-1})^{-1}$.Then, for all $k \geq 1$, we have

$$\lambda_k \leq \left(1-(2\varkappa)^{-1}\right)^k \sqrt{\tfrac{3}{2}}\,\lambda_0,$$

$$\lambda_k \leq \left[\tfrac{5}{2}\varkappa_\tau\left((2\varkappa)^{13n/(6k)}-1\right)\right]^{k/2} \sqrt{\tfrac{3}{2}\varkappa}\,\lambda_0.$$

From Gradient Norms to Function Values

- For self-concordant functions, we have

$$\frac{\lambda^2(x)}{2 + M\lambda(x)} \leq f(x) - f^* \leq \frac{\lambda^2(x)}{2 - M\lambda(x)}$$

  for any x such that $M\lambda(x) < 2$, where $\lambda(x) := \|\nabla f(x)\|_x^*$.
- In particular, if $M\lambda(x) \leq 1$, then

$$\tfrac{1}{3}\lambda^2(x) \leq f(x) - f^* \leq \lambda^2(x).$$

- In our methods, $M\lambda_k \leq 1$ for all $k \geq 0$. Thus,

$$\lambda_k^2 \leq (\tfrac{1}{3}\epsilon)\lambda_0^2 \quad \Longrightarrow \quad f(x_k) - f^* \leq \epsilon[f(x_0) - f^*].$$

Change of Notation

Direct QN update: For $G \in \mathbb{S}^n_{++}$, $\delta, \gamma \in \mathbb{R}^n$, define

$$\mathrm{Upd}(G, \delta, \gamma) := [\mathrm{Upd}^{-1}(G^{-1}, \delta, \gamma)]^{-1}.$$

Explicit formulas:

$$\mathrm{BFGS}(G, \delta, \gamma) = G - \frac{G\delta\delta^T G}{\langle G\delta, \delta\rangle} + \frac{\gamma\gamma^T}{\langle \gamma, \delta\rangle},$$

$$\mathrm{DFP}(G, \delta, \gamma) = G - \frac{G\delta\gamma^T + \gamma\delta^T G}{\langle \gamma, \delta\rangle} + \left(\frac{\langle G\delta, \delta\rangle}{\langle \gamma, \delta\rangle} + 1\right)\frac{\gamma\gamma^T}{\langle \gamma, \delta\rangle}.$$

Matrix-revealing form: For $G, A \in \mathbb{S}^n_{++}$ and $u \in \mathbb{R}^n$, define

$$\mathrm{Upd}(G, A, u) := \mathrm{Upd}(G, u, Au).$$

Classical QN update:

$$u = x_+ - x,\ A = \int_0^1 \nabla^2 f(x + tu)dt \implies Au = \nabla f(x_+) - \nabla f(x).$$

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Eigenvalue Property

Hereinafter, $A, G \in \mathbb{S}^n_{++}$, $u \in \mathbb{R}^n$, $\tau \in [0, 1]$ are arbitrary and

$$G_+ = \mathrm{Broyd}_\tau(G, A, u).$$

Eigenvalue property

For any $\xi, \eta \geq 1$, the following implication holds:

$$\xi^{-1} A \preceq G \preceq \eta A \quad \Longrightarrow \quad \xi^{-1} A \preceq G_+ \preceq \eta A.$$

Quality of Approximation

Directional measure of closeness

$$\nu(G, A, u) := \frac{\|(G - A)u\|^*_{G_+}}{\|u\|_G}.$$

Here $\|u\|_G := \langle Gu, u\rangle^{1/2}$, $\|s\|^*_{G_+} := \langle s, G_+^{-1}s\rangle^{1/2}$.

Note: If $u = x_+ - x = -G^{-1}\nabla f(x)$ and $A = \int_0^1 \nabla^2 f(x + tu)dt$, then

$$\nu(G, A, u) = \frac{\|\nabla f(x_+)\|^*_{G_+}}{\|\nabla f(x)\|^*_G}$$

because $Au = \nabla f(x_+) - \nabla f(x)$.

Our goal: Show that $\nu \to 0$ (and estimate the rate of convergence).

Potential Function

Augmented Log-Det Barrier

For $X, Y \succ 0$, define

$$\psi(X, Y) := -\ln\det Y + \ln\det X + \langle X^{-1}, Y - X\rangle,$$

where $\langle U, V\rangle := \operatorname{tr}(UV)$ is the Frobenius inner product.

Remarks:

- This is the Bregman divergence generated by $d(X) := -\ln\det X$:

$$\psi(X, Y) = d(Y) - d(X) - \langle \nabla d(X), Y - X\rangle \geq 0.$$

- First used in (Byrd and Nocedal, 1989) for the analysis of QN methods.

Key Result

Key result

If $\xi^{-1}A \preceq G \preceq \eta A$ for some $\xi, \eta \geq 1$, then

$$\psi(G_+, A) \leq \psi(G, A) - \frac{6}{13} \ln(1 + \delta\nu^2),$$

where $\delta := \frac{1}{1+\xi}(1 - \tau + \tau\frac{1}{\xi\eta})$ and $\nu := \nu(G, A, u)$.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Minimizing Quadratic Function

Problem

$$\min_{x \in \mathbb{R}^n} \left[f(x) := \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle \right],$$

where $A \in \mathbb{S}^n_{++}$ and $b \in \mathbb{R}^n$.

Main assumption: $\mu I \preceq A \preceq LI$ for some $\mu, L > 0$.

Convex Broyden Method

Given $x_0 \in \mathbb{R}^n$, set $G_0 = LI$, and iterate for $k \geq 0$:

1. Set $x_{k+1} = x_k - G_k^{-1} \nabla f(x_k)$.
2. Set $G_{k+1} = \mathrm{Broyd}_\tau(G_k, A, u_k)$ for $u_k := x_{k+1} - x_k$.

Accuracy measure: $\lambda_k = \|\nabla f(x_k)\|_A^*$.

Linear Convergence

Bounds on Hessian approximations: $A \preceq G_k \preceq \varkappa A, \quad \forall k \geq 0.$

Proof: Indeed, by construction

$$A \preceq G_0 = LI = \varkappa(\mu I) \preceq \varkappa A,$$

and each subsequent update preserves these bounds. □

Corollary (linear convergence): $\lambda_{k+1} \leq (1 - \varkappa^{-1})\lambda_k, \quad \forall k \geq 0.$

Proof: Recall that $u_k = x_{k+1} - x_k = -G_k^{-1}\nabla f(x_k)$. Hence,

$$\nabla f(x_{k+1}) = \nabla f(x_k) + Au_k = \nabla f(x_k) - AG_k^{-1}\nabla f(x_k).$$

Therefore,

$$\lambda_{k+1} \equiv \|\nabla f(x_{k+1})\|_A^* = \|(A^{-1} - G_k^{-1})\nabla f(x_k)\|_A^* \leq (1 - \varkappa^{-1})\lambda_k. \quad \square$$

Superlinear Sonvergence

$$\lambda_k \leq [\varkappa_\tau(\varkappa^{13n/(6k)} - 1)]^{k/2}\sqrt{\varkappa}\,\lambda_0, \qquad \forall k \geq 1,$$

where $\varkappa_\tau := 2(1 - \tau + \tau\varkappa^{-1})^{-1}$.

Proof: Using the key result, we get, for any $i \geq 0$,

$$\phi(\nu_i) := \tfrac{6}{13}\ln(1 + \varkappa_\tau \nu_i^2) \leq \psi_i - \psi_{i+1}, \qquad \nu_i = \frac{g_{i+1}}{g_i},$$

where $g_i := \|\nabla f(x_i)\|^*_{G_i}$ and $\psi_i := \psi(G_i, A) \geq 0$. Hence,

$$\sum_{i=0}^{k-1} \phi(\nu_i) \leq \psi_0 \leq n \ln \varkappa.$$

By convexity of $t \mapsto \ln(1 + e^t)$, it follows that

$$\frac{n \ln \varkappa}{k} \geq \frac{1}{k}\sum_{i=0}^{k-1} \phi(\nu_i) \geq \phi\Big(\Big[\prod_{i=0}^{k-1} \nu_i\Big]^{1/k}\Big) = \phi\Big(\Big[\frac{g_k}{g_0}\Big]^{1/k}\Big).$$

It remains to rearrange and connect g_k with λ_k. □

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Nonlinear Functions

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

Convex Broyden Method

Given $x_0 \in \mathbb{R}^n$, set $G_0 = LI$, and iterate for $k \geq 0$:

1. Set $x_{k+1} = x_k - G_k^{-1} \nabla f(x_k)$.
2. Set $G_{k+1} = \mathrm{Broyd}_\tau(G_k, J_k, u_k)$,
   where $u_k := x_{k+1} - x_k$, $J_k := \int_0^1 \nabla^2 f(x_k + t u_k) dt$.

Main difficulty (compared to quadratic case): J_k changes with k.

But locally all Hessians are close to each other:

For any $x, y \in \mathbb{R}^n$, $J := \int_0^1 \nabla^2 f(x + t(y - x)) dt$, $r := \|y - x\|_x$, $z \in \{x, y\}$,

$$(1 + Mr)^{-1} \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq (1 + Mr) \nabla^2 f(x),$$

$$(1 + \tfrac{1}{2} Mr)^{-1} \nabla^2 f(z) \preceq J \preceq (1 + \tfrac{1}{2} Mr) \nabla^2 f(z).$$

Proof Idea

Local gradient norm: $\lambda_k := \|\nabla f(x_k)\|^*_{x_k}$.

Denote $\xi_k := \exp(M \sum_{i=0}^{k-1} r_i)$ (≥ 1), where $r_k := \|u_k\|_{x_k}$.

For all $k \geq 1$, we have

$$\xi_k^{-1} \nabla^2 f(x_k) \preceq G_k \preceq (\xi_k \varkappa) \nabla^2 f(x_k), \qquad \xi_{k+1}^{-1} J_k \preceq G_k \preceq (\xi_{k+1} \varkappa) J_k,$$

$$\lambda_k \leq q^k \sqrt{\xi_k}\, \lambda_0,$$

$$\lambda_k \leq \left[\varkappa_k \left((\xi_{k+1}^{\xi_{k+1}} \varkappa)^{13n/(6k)} - 1 \right)\right]^{k/2} \sqrt{\xi_k \varkappa}\, \lambda_0,$$

for $q := \max\{1 - (\xi_k \varkappa)^{-1}, \xi_k - 1\}$, $\varkappa_k := (1 + \xi_k)(1 - \tau + \tau \xi_k^{-2} \varkappa^{-1})^{-1}$.

- For a quadratic function, we had $M = 0 \implies \xi_k \equiv 1$.
- **Goal:** Prove that $\xi_k \leq \xi$, $\forall k \geq 0$, assuming λ_0 is small enough.
- Suffices to prove by induction: $r_k \leq \xi_k \lambda_k \leq O(\rho^k \lambda_0)$ for $\rho \in (0, 1)$.

Outline

1. Introduction
   - Optimization problems
   - Gradient method
   - Newton's method
   - Quasi-Newton (QN) methods
   - Classical results on QN methods
2. Modern local analysis of QN methods
   - Problem formulation and assumptions
   - Final complexity bound
   - Proof technique: preliminaries
   - Two main properties of convex Broyden updates
   - Analysis for quadratic functions
   - Nonlinear functions
   - Conclusions

Conclusion

- We finally have some complexity bounds for QN methods.
- Theory confirms (well-known) superiority of BFGS over DFP.
- Complexity result for BFGS is very attractive:

$$\min\Big\{ \underbrace{\varkappa \ln O(\epsilon^{-1})}_{\substack{\text{Complexity of} \\ \text{Gradient method}}}, \underbrace{O\big(n \ln(2\varkappa)\big)}_{\substack{\text{Start of super.} \\ \text{convergence}}} + \ln O(\epsilon^{-1}) \Big\}.$$

Still many interesting open questions:

- Optimality of our results.
- Choice of initial matrix (line search?).
- Global complexity bounds.
- Limited-memory QN methods (L-BFGS).
- Application to Interior-Point methods.
- Composite optimization.
- Acceleration.

Thank you!

References I


C. G. Broyden, J. E. Dennis Jr, and J. Moré. On the local and superlinear convergence of quasi-Newton methods. *IMA Journal of Applied Mathematics*, 12(3):223–245, 1973.


R. Byrd and J. Nocedal. A tool for the analysis of quasi-Newton methods with application to unconstrained minimization. *SIAM Journal on Numerical Analysis*, 26(3):727–739, 1989.


A. R. Conn, N. I. Gould, and P. L. Toint. *Trust Region Methods*. SIAM, 2000.


W. Davidon. Variable metric method for minimization. Technical report 5990, Argonne National Laboratory, 1959.


R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *Computer Journal*, 6(2):163–168, 1963.


S. M. Goldfeld, R. E. Quandt, and H. F. Trotter. Maximization by quadratic hill-climbing. *Econometrica: Journal of the Econometric Society*:541–551, 1966.

References II


K. Levenberg. A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics*, 2(2):164–168, 1944.


D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.


Y. Nesterov and B. T. Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108:177–205, 2006.


M. J. D. Powell. Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In R. W. Cottle and C. E. Lemke, editors, *Nonlinear Programming, SIAM-AMS proceedings*, volume 9. American Mathematical Society, 1976.


A. Rodomanov and Y. Nesterov. Greedy Quasi-Newton Methods with Explicit Superlinear Convergence. *SIAM Journal on Optimization*, 31(1):785–811, 2021.

References III

A. Rodomanov and Y. Nesterov. New Results on Superlinear Convergence of Classical Quasi-Newton Methods. *Journal of Optimization Theory and Applications*, 188:744–769, 2021.

A. Rodomanov and Y. Nesterov. Rates of superlinear convergence for classical quasi-Newton methods. *Mathematical Programming*, 194:159–190, 2022.