Non-asymptotic Global Convergence Analysis of BFGS with the Armijo-Wolfe Line Search

Qiujiang Jin, Ruichen Jiang, Aryan Mokhtari

ECE Department, UT Austin

Neurips 2024

What is the problem of interest?

- Consider the general unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$

  - **Assumption 1:** $f(x)$ is strongly convex with $\mu > 0$.

What is the problem of interest?

- Consider the general unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$

  - **Assumption 1:** $f(x)$ is strongly convex with $\mu > 0$.
  - **Assumption 2:** The gradient $\nabla f(x)$ is Lipschitz continuous with $L > 0$.

What is the problem of interest?

- Consider the general unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$

  - **Assumption 1:** $f(x)$ is strongly convex with $\mu > 0$.
  - **Assumption 2:** The gradient $\nabla f(x)$ is Lipschitz continuous with $L > 0$.
  - **Assumption 3:** The Hessian $\nabla^2 f(x)$ is Lipschitz continuous with $M > 0$.

What is the problem of interest?

- Consider the general unconstrained minimization problem

$$\min_{x \in \mathbb{R}^d} \quad f(x),$$

  - **Assumption 1:** $f(x)$ is strongly convex with $\mu > 0$.
  - **Assumption 2:** The gradient $\nabla f(x)$ is Lipschitz continuous with $L > 0$.
  - **Assumption 3:** The Hessian $\nabla^2 f(x)$ is Lipschitz continuous with $M > 0$.
- **Goal:** Finding the global complexity of classic quasi-Newton methods for this setting

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up GD-type methods by approximating the function's curvature and using a preconditioner

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k)$$

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up GD-type methods by approximating the function's curvature and using a preconditioner

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k)$$

- When $B_k \approx \nabla^2 f(x_k)$ they mimic Newton's method

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up GD-type methods by approximating the function's curvature and using a preconditioner

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k)$$

- When $B_k \approx \nabla^2 f(x_k)$ they mimic Newton's method
- Only use gradient to construct $B_k \Rightarrow$ Still first-order methods

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up GD-type methods by approximating the function's curvature and using a preconditioner

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k)$$

- When $B_k \approx \nabla^2 f(x_k)$ they mimic Newton's method
- Only use gradient to construct $B_k \Rightarrow$ Still first-order methods
- Main ideas:
  - *Proximity condition*: Keep B_k and B_{k+1} close
  - *Secant condition*: $B_{k+1} s_k = y_k$ where $s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up GD-type methods by approximating the function's curvature and using a preconditioner

$$x_{k+1} = x_k - \eta_k B_k^{-1} \nabla f(x_k)$$

- When $B_k \approx \nabla^2 f(x_k)$ they mimic Newton's method
- Only use gradient to construct $B_k \Rightarrow$ Still first-order methods

- Main ideas:
  - *Proximity condition*: Keep B_k and B_{k+1} close
  - *Secant condition*: $B_{k+1} s_k = y_k$ where $s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$

$$\begin{aligned} B_{k+1} \; = \; & \operatorname{argmin} \; \|B - B_k\|_{\mathbf{V}} \\ & \text{s.t.} \;\; B\, s_k = y_k, \quad B \succeq \mathbf{0} \end{aligned}$$

BFGS quasi-Newton Method

- Focus on the BFGS quasi-Newton method:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k}.$$

BFGS quasi-Newton Method

- Focus on the BFGS quasi-Newton method:

$$B_{k+1} = B_k - \frac{B_k s_k s_k^\top B_k}{s_k^\top B_k s_k} + \frac{y_k y_k^\top}{s_k^\top y_k}.$$

- Define $H_k = B_k^{-1}$. Using Sherman-Morrison-Woodbury formula, we have

$$H_{k+1} = \left(I - \frac{s_k y_k^\top}{y_k^\top s_k}\right) H_k \left(I - \frac{y_k s_k^\top}{s_k^\top y_k}\right) + \frac{s_k s_k^\top}{y_k^\top s_k}.$$

State-of-the-art Results on Standard Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods: when $\|x_k - x^*\|$ is small,

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

State-of-the-art Results on Standard Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods: when $\|x_k - x^*\|$ is small,

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

  - Local superlinear rate [Broyden-Dennis-Moré'73][Dennis-Moré'74]
  - Global and superlinear rate with exact linesearch [Powell'71][Dixon'72]
  - Global and superlinear rate with inexact linesearch [Powell'76][Bryd-Nocedal-Yuan'87]
  - Many other works: [Griewank-Toint'82; Dennis-Martinez-Tapia'89; Yuan'91; Al-Baali'98; Li-Fukushima'99; Yabe-Ogasawara-Yoshino'07; M-Eisen-Ribeiro'18; Gao-Goldfarb'19]

State-of-the-art Results on Standard Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods: when $\|x_k - x^*\|$ is small,

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

  - Local superlinear rate [Broyden-Dennis-Moré'73][Dennis-Moré'74]
  - Global and superlinear rate with exact linesearch [Powell'71][Dixon'72]
  - Global and superlinear rate with inexact linesearch [Powell'76][Bryd-Nocedal-Yuan'87]
  - Many other works: [Griewank-Toint'82; Dennis-Martinez-Tapia'89; Yuan'91; Al-Baali'98; Li-Fukushima'99; Yabe-Ogasawara-Yoshino'07; M-Eisen-Ribeiro'18; Gao-Goldfarb'19]
- However, they are all asymptotic and fail to provide an explicit convergence rate

Recent Results on Quasi-Newton Methods

- Recent results show explicit non-asymptotic local superlinear rate for quasi-Newton methods

Recent Results on Quasi-Newton Methods

- Recent results show explicit non-asymptotic local superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed superlinear rates of the form $O((1/\sqrt{k})^k)$

Recent Results on Quasi-Newton Methods

- Recent results show explicit non-asymptotic local superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed superlinear rates of the form $O((1/\sqrt{k})^k)$

	cond. on $\|x_0 - x^*\|$	cond. on B_0	rate
[Jin-M'20]	$\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$	$B_0 \approx \nabla^2 f(x_0)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)^k$
[Rodomanov-Nesterov'20]	$\mathcal{O}\left(\frac{1}{d}\right)$	$\nabla^2 f(x) \preceq B_0 \preceq \kappa \nabla^2 f(x)$	$\mathcal{O}\left(\sqrt{\frac{d \ln \kappa}{k}}\right)^k$

Table: Definition $\kappa = L/\mu$

Recent Results on Quasi-Newton Methods

- Recent results show explicit non-asymptotic local superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed superlinear rates of the form $O((1/\sqrt{k})^k)$

	cond. on $\|x_0 - x^*\|$	cond. on B_0	rate
[Jin-M'20]	$\mathcal{O}\left(\frac{1}{\sqrt{d}}\right)$	$B_0 \approx \nabla^2 f(x_0)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)^k$
[Rodomanov-Nesterov'20]	$\mathcal{O}\left(\frac{1}{d}\right)$	$\nabla^2 f(x) \preceq B_0 \preceq \kappa \nabla^2 f(x)$	$\mathcal{O}\left(\sqrt{\frac{d \ln \kappa}{k}}\right)^k$

Table: Definition $\kappa = L/\mu$

- These results are only local, it is unclear how to extend them into global guarantees
  $\Rightarrow$ The condition on B_0 may not hold when $\|x_0 - x^*\|$ becomes small
- Moreover, there is no global result matching the linear rate of GD

Contributions

- One of the first global non-asymptotic analysis of classic quasi-Newton methods
  - Arbitrary initial point $x_0 \in \mathbb{R}^d$ and initial Hessian approximation $B_0 \in \mathbb{S}^d_{++}$

Contributions

- One of the first global non-asymptotic analysis of classic quasi-Newton methods
  - Arbitrary initial point $x_0 \in \mathbb{R}^d$ and initial Hessian approximation $B_0 \in \mathbb{S}^d_{++}$

- Focus on the Armijo-Wolfe Line Search scheme: if $d_k = -B_k^{-1} \nabla f(x_k)$,

$$f(x_k + \eta_k d_k) \leq f(x_k) + \alpha \eta_k \nabla f(x_k)^\top d_k,$$

$$\nabla f(x_k + \eta_k d_k)^\top d_k \geq \beta \nabla f(x_k)^\top d_k,$$

where α and β satisfy $0 < \alpha < \beta < 1$ and $0 < \alpha < \frac{1}{2}$.

Summary of Results for BFGS with Armijo-Wolfe LS

Matrix	Convergence Phase	Convergence Rate	Starting moment
B_0	Linear phase	$\left(1-\frac{1}{\kappa}\right)^k$	$\Psi(\bar{B}_0)$
B_0	Superlinear phase	$\left(\frac{\Psi(\tilde{B}_0)+C_0\Psi(\bar{B}_0)+C_0\kappa}{k}\right)^k$	$\Psi(\tilde{B}_0)+C_0\Psi(\bar{B}_0)+C_0\kappa$
LI	Linear phase	$\left(1-\frac{1}{\kappa}\right)^k$	1
LI	Superlinear phase	$\left(\frac{d\kappa+C_0\kappa}{k}\right)^k$	$d\kappa+C_0\kappa$
μI	Linear phase	$\left(1-\frac{1}{\kappa}\right)^k$	$d\log\kappa$
μI	Superlinear phase	$\left(\frac{(1+C_0)d\log\kappa+C_0\kappa}{k}\right)^k$	$(1+C_0)d\log\kappa+C_0\kappa$

- Here $C_0 := \frac{M}{\mu^{\frac{3}{2}}}\sqrt{2(f(x_0)-f(x_*))}$ and $\Psi(A) := \mathbf{Tr}(A) - \log \mathbf{Det}(A) - d$

Notation and Definitions

- Introduce a weight matrix $P \in \mathbb{S}^d_{++}$ and define

$$\hat{g}_k = P^{-\frac{1}{2}} g_k, \qquad \hat{y}_k = P^{-\frac{1}{2}} y_k, \qquad \hat{s}_k = P^{\frac{1}{2}} s_k, \qquad \hat{B}_k = P^{-\frac{1}{2}} B_k P^{-\frac{1}{2}}.$$

Notation and Definitions

- Introduce a weight matrix $P \in \mathbb{S}^d_{++}$ and define

$$\hat{g}_k = P^{-\frac{1}{2}} g_k, \qquad \hat{y}_k = P^{-\frac{1}{2}} y_k, \qquad \hat{s}_k = P^{\frac{1}{2}} s_k, \qquad \hat{B}_k = P^{-\frac{1}{2}} B_k P^{-\frac{1}{2}}.$$

- The weighted BFGS update still holds $\hat{B}_{k+1} = \hat{B}_k - \frac{\hat{B}_k \hat{s}_k \hat{s}_k^\top \hat{B}_k}{\hat{s}_k^\top \hat{B}_k \hat{s}_k} + \frac{\hat{y}_k \hat{y}_k^\top}{\hat{s}_k^\top \hat{y}_k}$.

Notation and Definitions

- Introduce a weight matrix $P \in \mathbb{S}^d_{++}$ and define

$$\hat{g}_k = P^{-\frac{1}{2}} g_k, \qquad \hat{y}_k = P^{-\frac{1}{2}} y_k, \qquad \hat{s}_k = P^{\frac{1}{2}} s_k, \qquad \hat{B}_k = P^{-\frac{1}{2}} B_k P^{-\frac{1}{2}}.$$

- The weighted BFGS update still holds $\quad \hat{B}_{k+1} = \hat{B}_k - \frac{\hat{B}_k \hat{s}_k \hat{s}_k^\top \hat{B}_k}{\hat{s}_k^\top \hat{B}_k \hat{s}_k} + \frac{\hat{y}_k \hat{y}_k^\top}{\hat{s}_k^\top \hat{y}_k}.$
- P plays critical roles in the proof of non-asymptotic convergence rates.
  - $\Rightarrow$ Choose $P = LI$ to prove the linear convergence rates.
  - $\Rightarrow$ Choose $P = \nabla^2 f(x_*)$ to prove the superlinear convergence rates.

Notation and Definitions

- Introduce a weight matrix $P \in \mathbb{S}^d_{++}$ and define

$$\hat{g}_k = P^{-\frac{1}{2}} g_k, \qquad \hat{y}_k = P^{-\frac{1}{2}} y_k, \qquad \hat{s}_k = P^{\frac{1}{2}} s_k, \qquad \hat{B}_k = P^{-\frac{1}{2}} B_k P^{-\frac{1}{2}}.$$

- The weighted BFGS update still holds $\hat{B}_{k+1} = \hat{B}_k - \frac{\hat{B}_k \hat{s}_k \hat{s}_k^\top \hat{B}_k}{\hat{s}_k^\top \hat{B}_k \hat{s}_k} + \frac{\hat{y}_k \hat{y}_k^\top}{\hat{s}_k^\top \hat{y}_k}$.
- P plays critical roles in the proof of non-asymptotic convergence rates.
  - $\Rightarrow$ Choose $P = LI$ to prove the linear convergence rates.
  - $\Rightarrow$ Choose $P = \nabla^2 f(x_*)$ to prove the superlinear convergence rates.

- Define the following terms

$$\hat{p}_k := \frac{f(x_k) - f(x_{k+1})}{-\hat{g}_k^\top \hat{s}_k}, \quad \hat{q}_k := \frac{\|\hat{g}_k\|^2}{f(x_k) - f(x_*)}, \quad \hat{m}_k := \frac{\hat{y}_k^\top \hat{s}_k}{\|\hat{s}_k\|^2}, \quad \hat{n}_k = \frac{\hat{y}_k^\top \hat{s}_k}{-\hat{g}_k^\top \hat{s}_k}.$$

$$\cos(\theta_k) := \frac{g_k^\top B_k^{-1} g_k}{\|g_k\| \|B_k^{-1} g_k\|}$$

Key Lemma

Lemma: [Jin-Jiang-M, 2024]

Let $\{x_k\}_{k\geq 0}$ be the iterates generated by the BFGS method with AW line search. Given a weight matrix $P \in \mathbb{S}^d_{++}$, for any $k \geq 1$, we have

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \left(\prod_{i=0}^{k-1} \hat{p}_i \hat{q}_i \hat{n}_i \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}\right)^{\frac{1}{k}}\right)^k.$$

- Fundamental framework in the whole convergence analysis.
- Used for the proof of both linear and superlinear convergence rates.
- Need to lower bound the following three products

$$\prod_{i=0}^{k-1} \hat{p}_i, \quad \prod_{i=0}^{k-1} \hat{q}_i, \quad \prod_{i=0}^{k-1} \hat{n}_i, \quad \prod_{i=0}^{k-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}$$

Lower bounds on $\hat{p}_k$ and $\hat{n}_k$

Lemma: [Jin-Jiang-M, 2024]

For the BFGS method with Armijo-Wolfe line search, we have

$$\frac{f(x_k) - f(x_{k+1})}{-g_k^\top s_k} \geq \alpha, \qquad \frac{y_k^\top s_k}{-g_k^\top s_k} \geq 1 - \beta, \qquad \textit{and} \qquad f(x_{k+1}) \leq f(x_k).$$

Lower bounds on $\hat{p}_k$ and $\hat{n}_k$

Lemma: [Jin-Jiang-M, 2024]

For the BFGS method with Armijo-Wolfe line search, we have

$$\frac{f(x_k) - f(x_{k+1})}{-g_k^\top s_k} \geq \alpha, \qquad \frac{y_k^\top s_k}{-g_k^\top s_k} \geq 1 - \beta, \qquad \textit{and} \qquad f(x_{k+1}) \leq f(x_k).$$

Given $\hat{p}_k := \frac{f(x_k)-f(x_{k+1})}{-\hat{g}_k^\top \hat{s}_k}$ and $\hat{n}_k = \frac{\hat{y}_k^\top \hat{s}_k}{-\hat{g}_k^\top \hat{s}_k}$

Lemma: [Jin-Jiang-M, 2024]

Then, for any $k \geq 0$ *and any weight matrix* $P \in \mathbb{S}^d_{++}$

$$\hat{p}_k \geq \alpha, \qquad \hat{n}_k \geq 1 - \beta$$

Lower bounds on $\hat{q}_k$

- Define C_k as the measurement of distance between x_k and x_*

$$C_k := \frac{M}{\mu^{\frac{3}{2}}}\sqrt{2(f(x_k) - f(x_*))}.$$

Lemma: [Jin-Jiang-M, 2024]

Recall the definition $\hat{q}_k = \frac{\|\hat{g}_k\|^2}{f(x_k)-f(x_*)}$. Then we have the following results:

(a) *If we choose $P = LI$, then $\hat{q}_k \geq 2/\kappa$.*

(b) *If we choose $P = \nabla^2 f(x_*)$, then $\hat{q}_k \geq 2/(1 + C_k)^2$.*

- Depends on the choice of the weight matrix $P \in \mathbb{S}^d_{++}$.

Lower bounds on $\frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}$

- Define the trace and log-determinant potential function for any $A \in \mathbb{S}^d_{++}$ as

$$\Psi(A) := \mathbf{Tr}(A) - \log \mathbf{Det}(A) - d.$$

- The Bregman divergence between matrix A and the identity matrix I.
- $\Psi(A) \geq 0$ and $\Psi(A) = 0$ holds if and only if $A = I$.

Lemma: [Jin-Jiang-M, 2024]

For the BFGS method, we have that

(a) *If* $P = LI$, *then* $\prod_{i=0}^{k-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq e^{-\Psi(\bar{B}_0)}$.

(b) *If* $P = \nabla^2 f(x_*)$, *then* $\prod_{i=0}^{k-1} \frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i} \geq e^{-\Psi(\hat{B}_0) - \sum_{i=0}^{k-1} C_i}$.

Global Linear Convergence Rates

- For the global linear results we use $P = LI$, hence $\bar{B}_k = (1/L)B_k$

Theorem: [Jin-Jiang-M, 2024]

Consider BFGS with Armijo-Wolfe line search. For any initial point $x_0 \in \mathbb{R}^d$ and any initial Hessian approximation $B_0 \in \mathbb{S}^d_{++}$, the following global convergence rates hold,

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - e^{-\frac{\Psi(\bar{B}_0)}{k}} \frac{2\alpha(1-\beta)}{\kappa}\right)^k,$$

Global Linear Convergence Rates

- For the global linear results we use $P = LI$, hence $\bar{B}_k = (1/L)B_k$

Theorem: [Jin-Jiang-M, 2024]

Consider BFGS with Armijo-Wolfe line search. For any initial point $x_0 \in \mathbb{R}^d$ and any initial Hessian approximation $B_0 \in \mathbb{S}^d_{++}$, the following global convergence rates hold,

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - e^{-\frac{\Psi(\bar{B}_0)}{k}} \frac{2\alpha(1-\beta)}{\kappa}\right)^k,$$

Special Cases:

- $B_0 = LI$: For all $k \geq 1$ $\quad \frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)} \leq \left(1 - \frac{2\alpha(1-\beta)}{\kappa}\right)^k.$
- $B_0 = \mu I$: For all $k \geq d \log \kappa$ $\quad \frac{f(x_k)-f(x_*)}{f(x_0)-f(x_*)} \leq \left(1 - \frac{2\alpha(1-\beta)}{3\kappa}\right)^k.$

Condition Number Independent Linear Rate

- Replace the bounds for $\frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}$ and $\hat{q}_i$ by the ones obtained using $P = \nabla^2 f(x_*)$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - 2\alpha(1-\beta)e^{-\frac{\Psi(\tilde{B}_0) + 3\sum_{i=0}^{k-1} C_i}{k}}\right)^k, \quad \forall k \geq 1.$$

- Now by bounding $\sum_{i=0}^{k-1} C_i$ using the previous linear result, we obtain the following

Condition Number Independent Linear Rate

- Replace the bounds for $\frac{\cos^2(\hat{\theta}_i)}{\hat{m}_i}$ and $\hat{q}_i$ by the ones obtained using $P = \nabla^2 f(x_*)$

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - 2\alpha(1-\beta)e^{-\frac{\Psi(\tilde{B}_0)+3\sum_{i=0}^{k-1} C_i}{k}}\right)^k, \quad \forall k \geq 1.$$

- Now by bounding $\sum_{i=0}^{k-1} C_i$ using the previous linear result, we obtain the following

Theorem: [Jin-Jiang-M, 2024]

Consider BFGS with Armijo-Wolfe LS. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, if $k \geq \Psi(\tilde{B}_0) + 3C_0\Psi(\bar{B}_0) + \frac{9}{\alpha(1-\beta)}C_0\kappa$ we have

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \frac{2\alpha(1-\beta)}{3}\right)^k.$$

- If we set $B_0 = LI$, the rate holds for $k \geq d\kappa + \frac{9}{\alpha(1-\beta)}C_0\kappa$,
- If we set $B_0 = \mu I$, the rate holds for $k \geq (1 + 3C_0)d\log\kappa + \frac{9}{\alpha(1-\beta)}C_0\kappa$.

Requirement for SuperLinear Rate

- To achieve a superlinear result we need tighter bounds: $\hat{p}_k \geq \alpha$ and $\hat{n}_k \geq 1 - \beta$
- We show that if $\eta = 1$ satisfies AW conditions, tighter bounds are achievable.

Lemma: [Jin-Jiang-M, 2024]

If $\eta_k = 1$ satisfies the conditions for Armijo-Wolfe LS, then we have

$$\hat{p}_k \geq 1 - \frac{1 + C_k}{2}, \qquad \hat{n}_k \geq \frac{1}{(1 + C_k)}.$$

Requirement for SuperLinear Rate

- To achieve a superlinear result we need tighter bounds: $\hat{p}_k \geq \alpha$ and $\hat{n}_k \geq 1 - \beta$
- We show that if $\eta = 1$ satisfies AW conditions, tighter bounds are achievable.

Lemma: [Jin-Jiang-M, 2024]

If $\eta_k = 1$ satisfies the conditions for Armijo-Wolfe LS, then we have

$$\hat{p}_k \geq 1 - \frac{1 + C_k}{2}, \qquad \hat{n}_k \geq \frac{1}{(1 + C_k)}.$$

Lemma: (Informal) [Jin-Jiang-M, 2024]

For $k \geq \max\left\{\Psi(\bar{B}_0), \frac{3\kappa}{\alpha(1-\beta)} \log \frac{C_0}{\delta_1}\right\}$, the number of time indices for which $\eta = 1$ does not satisfy the AWLS conditions is upper bounded.

Global Superlinear Rate

Theorem: [Jin-Jiang-M, 2024]

Consider BFGS with Armijo-Wolfe LS. For any $x_0 \in \mathbb{R}^d$ *and any* $B_0 \in \mathbb{S}^d_{++}$*, we have*

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{\Psi(\tilde{B}_0) + (1 + C_0)\Psi(\bar{B}_0) + (1 + C_0)\kappa}{k}\right)^k,$$

Global Superlinear Rate

Theorem: [Jin-Jiang-M, 2024]

Consider BFGS with Armijo-Wolfe LS. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, we have

$$\frac{f(x_k) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{\Psi(\tilde{B}_0) + (1 + C_0)\Psi(\bar{B}_0) + (1 + C_0)\kappa}{k}\right)^k,$$

- If $B_0 = LI$ BFGS achieves a rate of $\mathcal{O}((\frac{d\kappa + C_0\kappa}{k})^k)$
- If $B_0 = \mu I$ BFGS achieves a rate of $\mathcal{O}((\frac{C_0 d \log \kappa + C_0\kappa}{k})^k)$.
- Hence, the superlinear result for $B_0 = \mu I$ outperforms the rate for $B_0 = LI$ when $C_0 \log \kappa \ll \kappa$.

Numerical Experiments

- We focus on a hard cubic objective function, i.e.,

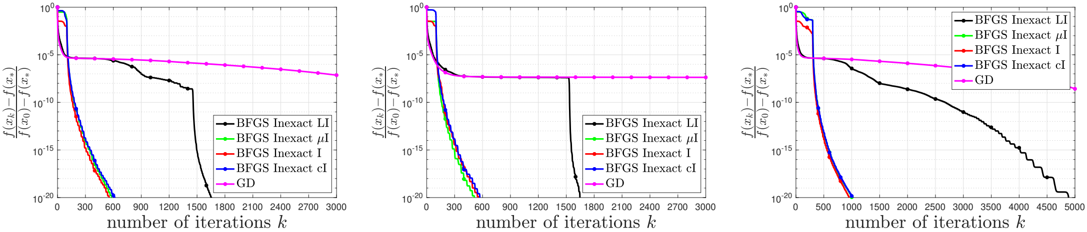
$$f(x) = \frac{\alpha}{12}\left(\sum_{i=1}^{d-1} g(v_i^\top x - v_{i+1}^\top x) - \beta v_1^\top x\right) + \frac{\lambda}{2}\|x\|^2,$$

  and $g : \mathbb{R} \to \mathbb{R}$ is defined as

$$g(w) = \begin{cases} \frac{1}{3}|w|^3 & |w| \leq \Delta, \\ \Delta w^2 - \Delta^2|w| + \frac{1}{3}\Delta^3 & |w| > \Delta, \end{cases}$$

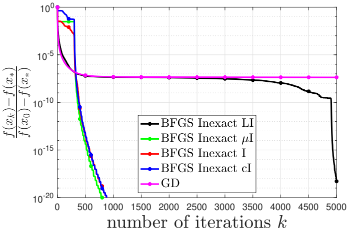
  where $\alpha, \beta, \lambda, \Delta \in \mathbb{R}$ are hyper-parameters and $\{v_i\}_{i=1}^n$ are standard orthogonal unit vectors in $\mathbb{R}^d$.

- This hard cubic function is used to establish a lower bound for second-order methods.

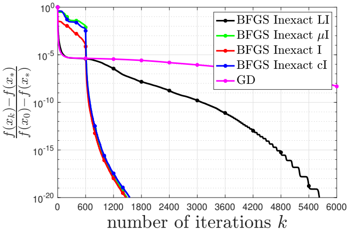


(a) $d = 100, \kappa = 100$. (b) $d = 100, \kappa = 1000$. (c) $d = 300, \kappa = 100$.

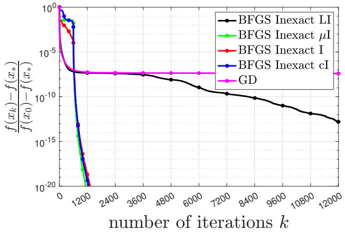
Figure: Convergence rates of BFGS with $B_0 = LI$, $B_0 = \mu I$, $B_0 = I$, $B_0 = cI$ and gradient descent to hard cubic objective function. $c = \frac{s^\top y}{\|s\|^2}$, with $s = x_2 - x_1$, $y = \nabla f(x_2) - \nabla f(x_1)$, and x_1, x_2 as two randomly generated vectors.



(a) $d = 300, \kappa = 1000$.



(b) $d = 600, \kappa = 100$.



(c) $d = 600, \kappa = 1000$.

Figure: Convergence rates of BFGS with $B_0 = LI$, $B_0 = \mu I$, $B_0 = I$, $B_0 = cI$ and gradient descent to hard cubic objective function. $c = \frac{s^\top y}{\|s\|^2}$, with $s = x_2 - x_1$, $y = \nabla f(x_2) - \nabla f(x_1)$, and x_1, x_2 as two randomly generated vectors.

Discussions on the line search complexity

- We proposed a Log Bisection Algorithm for finding a stepsize
- We showed when we run BFGS for N iterations:

  $\Rightarrow$ then the total number of function and gradient evaluations is

$$\mathcal{O}(N \max\{\log d, \log \kappa, \log C_0\})$$

- With more refine analysis, we can show that if $N = \Omega(\Psi(\tilde{B}_0) + (\Psi(\bar{B}_0) + \frac{3}{\alpha(1-\beta)}\kappa) C_0)$

  $\Rightarrow$ then the total line search complexity becomes $\mathcal{O}(N)$.