Affine-Invariant Global Non-Asymptotic Convergence Analysis of BFGS under Self-Concordance

Qiujiang Jin, Aryan Mokhtari

ECE Department, UT Austin

Neurips 2025

What is the problem of interest?

- Consider the general unconstrained minimization problem with optimal solution x_*

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

What is the problem of interest?

- Consider the general unconstrained minimization problem with optimal solution x_*

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

  - **Assumption 1:** $f(x)$ is twice differentiable and strictly convex.

What is the problem of interest?

- Consider the general unconstrained minimization problem with optimal solution x_*

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

  - **Assumption 1:** $f(x)$ is twice differentiable and strictly convex.
  - **Assumption 2:** $f(x)$ is strongly self-concordant with parameter $M > 0$, i.e.,

$$\nabla^2 f(x) - \nabla^2 f(y) \preceq M\|x - y\|_z \nabla^2 f(y), \qquad \forall x, y, z \in \mathbb{R}^d. \tag{1}$$

What is the problem of interest?

- Consider the general unconstrained minimization problem with optimal solution x_*

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

  - **Assumption 1:** $f(x)$ is twice differentiable and strictly convex.
  - **Assumption 2:** $f(x)$ is strongly self-concordant with parameter $M > 0$, i.e.,

$$\nabla^2 f(x) - \nabla^2 f(y) \preceq M\|x - y\|_z \nabla^2 f(y), \qquad \forall x, y, z \in \mathbb{R}^d. \tag{1}$$

- **Goal:** Finding the global complexity of classic quasi-Newton methods for this self-concordance setting.

Quasi-Newton Methods

- Quasi-Newton (QN) methods aim at speeding up first-order 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 first-order 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 first-order 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 first-order 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

- Focus on the BFGS quasi-Newton method:

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

Quasi-Newton Methods

- 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

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

Quasi-Newton Methods

- 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

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

- The computational cost per iteration is $\mathcal{O}(d^2)$

Quasi-Newton Methods

- 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

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

- The computational cost per iteration is $\mathcal{O}(d^2)$
- Focus on the weak Wolfe line search step size η_t with $d_t = -H_t \nabla f(x_t)$,

$$f(x_t + \eta_t d_t) \leq f(x_t) + \alpha \eta_t \nabla f(x_t)^\top d_t, \tag{2}$$

$$\nabla f(x_t + \eta_t d_t)^\top d_t \geq \beta \nabla f(x_t)^\top d_t, \tag{3}$$

where $0 < \alpha < \beta < 1$ and $0 < \alpha < 1/2$.

Quasi-Newton Methods

- The iterates of BFGS quasi-Newton method are affine invariant, i.e.,

Quasi-Newton Methods

- The iterates of BFGS quasi-Newton method are affine invariant, i.e.,

Theorem: [Jin-Mokhtari, 2025]

Let the iterations $\{x_t\}_{t=0}^{+\infty}$ be generated by the BFGS algorithm applied to the objective function $f(x)$. Consider the iterates $\{\dot{x}_t\}_{t=0}^{+\infty}$ produced by applying BFGS to the transformed function $\phi(x) = f(Ax)$, where $A \in \mathbb{R}^{d\times d}$ is a non-singular matrix. Assume that the initializations satisfy $\dot{x}_0 = A^{-1}x_0$ and $\dot{B}_0 = A^\top B_0 A$. Then, for any $t \geq 0$, the following relationships hold: $\dot{x}_t = A^{-1}x_t$, $\dot{B}_t = A^\top B_t A$ and $\phi(\dot{x}_t) = f(x_t)$.

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t\to\infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t\to\infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

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

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t\to\infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

- Recent results show explicit non-asymptotic superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed local superlinear rates of the form $O((1/\sqrt{t})^t)$

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t\to\infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

- Recent results show explicit non-asymptotic superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed local superlinear rates of the form $O((1/\sqrt{t})^t)$
- Jin-M'24a proved global non-asymptotic superlinear rate with exact linesearch

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t \to \infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

- Recent results show explicit non-asymptotic superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed local superlinear rates of the form $O((1/\sqrt{t})^t)$
- Jin-M'24a proved global non-asymptotic superlinear rate with exact linesearch
- Jin-M'24b and Rodomanov-Nesterov'24 proved global non-asymptotic superlinear rate with inexact linesearch

Resent Results on Quasi-Newton Methods

- Classic results have shown asymptotic local superlinear convergence for QN methods

$$\lim_{t\to\infty} \frac{\|x_{t+1} - x_*\|}{\|x_t - x_*\|} = 0$$

- Recent results show explicit non-asymptotic superlinear rate for quasi-Newton methods
- Rodomanov-Nesterov'20 and Jin-M'20 concurrently but using different Lyapunov functions showed local superlinear rates of the form $O((1/\sqrt{t})^t)$
- Jin-M'24a proved global non-asymptotic superlinear rate with exact linesearch
- Jin-M'24b and Rodomanov-Nesterov'24 proved global non-asymptotic superlinear rate with inexact linesearch
- However, all these results require strong convexity and are not affine invariant.

Contributions

- We establish first global non-asymptotic linear and superlinear convergence rates for BFGS without requiring strong convexity or Lipschitz continuity of the gradient or Hessian.

Contributions

- We establish first global non-asymptotic linear and superlinear convergence rates for BFGS without requiring strong convexity or Lipschitz continuity of the gradient or Hessian.
- We derive explicit convergence rates for the BFGS method that are affine invariant and consistent with the affine invariance property of the BFGS method.

Notation and Definitions

- For any $A \in \mathbb{S}^d_{++}$, we define $\Psi(A)$ as $\Psi(A) := \textbf{Trace}(A) - d - \log \textbf{Det}(A)$.

Notation and Definitions

- For any $A \in \mathbb{S}^d_{++}$, we define $\Psi(A)$ as $\Psi(A) := \textbf{Trace}(A) - d - \log \textbf{Det}(A)$.
- Introduce the function ω as $\omega(x) := x - \log(x+1)$ and ω^{-1} as its inverse function

Notation and Definitions

- For any $A \in \mathbb{S}^{d}_{++}$, we define $\Psi(A)$ as $\Psi(A) := \textbf{Trace}(A) - d - \log \textbf{Det}(A)$.
- Introduce the function ω as $\omega(x) := x - \log(x+1)$ and ω^{-1} as its inverse function
- Define the sequences $\{C_t\}_{t=0}^{+\infty}$ and $\{D_t\}_{t=0}^{+\infty}$ as

$$C_t := f(x_t) - f(x_*), \qquad D_t := 2\omega^{-1}\left(M^2 C_t/4\right).$$

Notation and Definitions

- For any $A \in \mathbb{S}^d_{++}$, we define $\Psi(A)$ as $\Psi(A) := \textbf{Trace}(A) - d - \log \textbf{Det}(A)$.
- Introduce the function ω as $\omega(x) := x - \log(x+1)$ and ω^{-1} as its inverse function
- Define the sequences $\{C_t\}_{t=0}^{+\infty}$ and $\{D_t\}_{t=0}^{+\infty}$ as

$$C_t := f(x_t) - f(x_*), \qquad D_t := 2\omega^{-1}\left(M^2 C_t/4\right).$$

- Define the following weighted versions of the initial Hessian approximation matrix B_0,

$$\bar{B}_0 = \frac{\nabla^2 f(x_*)^{-\frac{1}{2}} B_0 \nabla^2 f(x_*)^{-\frac{1}{2}}}{1 + D_0}, \qquad \tilde{B}_0 = \nabla^2 f(x_*)^{-\frac{1}{2}} B_0 \nabla^2 f(x_*)^{-\frac{1}{2}}.$$

Global Linear Convergence Rates

Theorem: [Jin-Mokhtari, 2025]

Consider BFGS with weak 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,

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \frac{\alpha(1-\beta)e^{-\frac{\Psi(\bar{B}_0)}{t}}}{(1+D_0)^2}\right)^t.$$

Moreover, when $t \geq \Psi(\bar{B}_0)$, we obtain that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} \leq \left(1 - \frac{\alpha(1-\beta)}{3(1+D_0)^2}\right)^t.$$

Global Linear Convergence Rates

- We proceed to present an improved version of the result of global linear convergence, showing that after a sufficient number of iterations, the linear rate of BFGS becomes independent of D_0 and B_0.

Global Linear Convergence Rates

- We proceed to present an improved version of the result of global linear convergence, showing that after a sufficient number of iterations, the linear rate of BFGS becomes independent of D_0 and B_0.

Theorem: [Jin-Mokhtari, 2025]

Consider BFGS with weak Wolfe LS. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, if $t \geq \Psi(\tilde{B}_0) + 3D_0(\Psi(\bar{B}_0) + \frac{3(1+D_0)^2}{\alpha(1-\beta)})$ we have

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

Requirement for SuperLinear Rate

- To achieve a superlinear result, we need establish under what conditions step size $\eta_t = 1$ is admissible.

Lemma: (Informal) [Jin-Mokhtari, 2025]

There exists constants $0 < \delta_1 < \delta_2 < 1 < \delta_3$. If $C_t \leq \delta_1$ and $\delta_2 \leq \rho_t \leq \delta_3$, then $\eta_t = 1$ satisfies weak Wolfe line search conditions.

Requirement for SuperLinear Rate

- To achieve a superlinear result, we need establish under what conditions step size $\eta_t = 1$ is admissible.

Lemma: (Informal) [Jin-Mokhtari, 2025]

There exists constants $0 < \delta_1 < \delta_2 < 1 < \delta_3$. If $C_t \leq \delta_1$ and $\delta_2 \leq \rho_t \leq \delta_3$, then $\eta_t = 1$ satisfies weak Wolfe line search conditions.

- Based on this, we show that the size of the set of indices where unit step size didn't satisfy the weak Wolfe conditions is limited.

Lemma: (Informal) [Jin-Mokhtari, 2025]

For $t \geq \max\left\{\Psi(\bar{B}_0), \frac{3(1+D_0)^2}{\alpha(1-\beta)} \log \frac{C_0}{\delta_1}\right\}$, the number of time indices for which $\eta = 1$ does not satisfy the weak Wolfe conditions is upper bounded.

Global Superlinear Rate

Theorem: [Jin-Mokhtari, 2025]

Consider BFGS with weak Wolfe line search. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, we have that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{\Psi(\tilde{B}_0) + D_0\Psi(\bar{B}_0) + (1 + D_0)^2}{t}\right)^t,$$

Global Superlinear Rate

Theorem: [Jin-Mokhtari, 2025]

Consider BFGS with weak Wolfe line search. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, we have that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{\Psi(\tilde{B}_0) + D_0\Psi(\bar{B}_0) + (1 + D_0)^2}{t}\right)^t,$$

- First non-asymptotic global superlinear convergence rate of a quasi-Newton method without the assumption of strong convexity.

Global Superlinear Rate

Theorem: [Jin-Mokhtari, 2025]

Consider BFGS with weak Wolfe line search. For any $x_0 \in \mathbb{R}^d$ and any $B_0 \in \mathbb{S}^d_{++}$, we have that

$$\frac{f(x_t) - f(x_*)}{f(x_0) - f(x_*)} = \mathcal{O}\left(\frac{\Psi(\tilde{B}_0) + D_0 \Psi(\bar{B}_0) + (1 + D_0)^2}{t}\right)^t,$$

- First non-asymptotic global superlinear convergence rate of a quasi-Newton method without the assumption of strong convexity.
- Both linear and superlinear convergence rates are affine invariant.

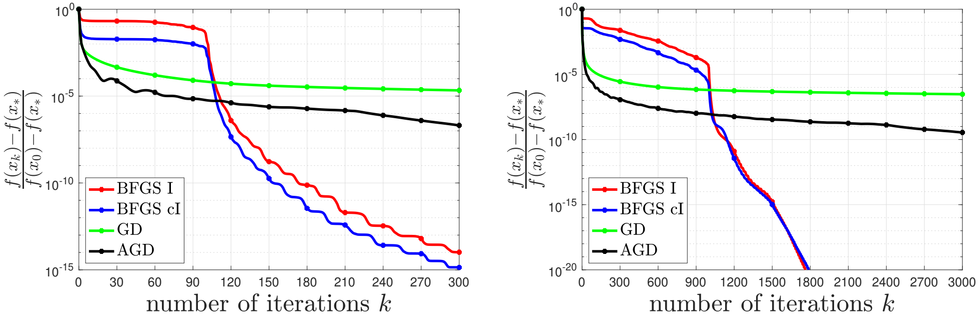
- We focus on a hard cubic objective function, i.e.,

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

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 $\mathbb{R}^d$.



(a) $d = 100$.   (b) $d = 1000$.

Figure: Convergence rates of BFGS with different B_0, gradient descent and accelerated gradient descent for solving the hard cubic function with different dimensions.

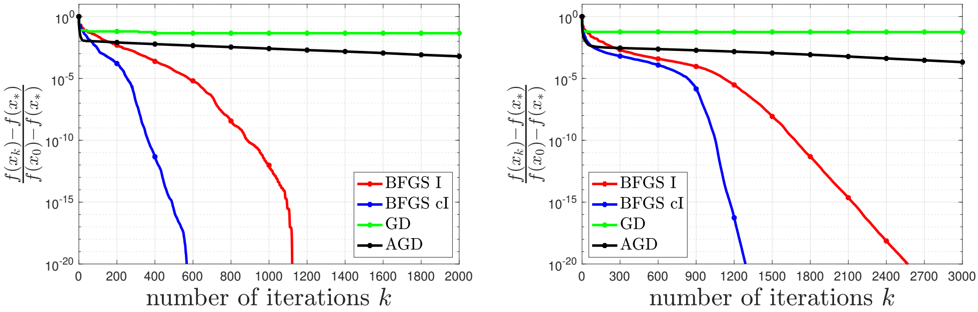
Numerical Experiments

- The second loss function is the logistic regression:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} \ln\left(1 + e^{-y_i z_i^\top x}\right),$$

where $\{z_i\}_{i=1}^N$ are the data points and $\{y_i\}_{i=1}^N$ are their corresponding labels.

Numerical Experiments



(a) $d = 100$.   (b) $d = 1000$.

Figure: Convergence rates of BFGS with different B_0, gradient descent and accelerated gradient descent for solving the logistic regression function with different dimensions.

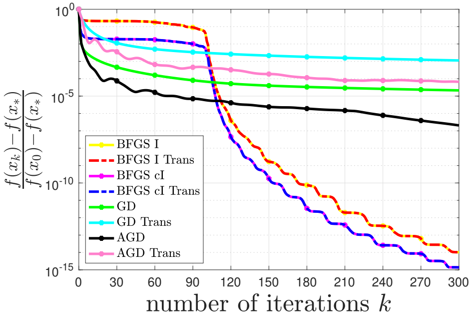
Numerical Experiments

- We compare the performance of BFGS, GD, and AGD under a transformation matrix A chosen to be a non-singular ill-conditioned matrix.

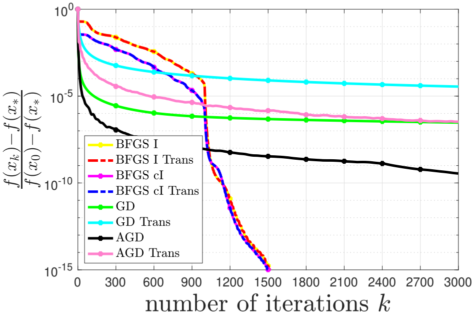
Numerical Experiments

- We compare the performance of BFGS, GD, and AGD under a transformation matrix A chosen to be a non-singular ill-conditioned matrix.
- We observe that the convergence trajectory of BFGS with this transformation is identical to that of the vanilla BFGS method, consistent with the affine invariance of quasi-Newton methods.

Numerical Experiments



(a) $d = 100$.



(b) $d = 1000$.

Figure: Convergence rates of BFGS with different B_0, gradient descent and accelerated gradient descent for solving the hard cubic function with transformation matrix A.