# Sharpened Quasi-Newton Methods: Faster Superlinear Rate and Larger Local Convergence Neighborhood

**Qiujiang Jin**[1], **Alec Koppel**[2], **Ketan Rajawat**[3], **Aryan Mokhtari**[1]




1. The University of Texas at Austin
2. Amazon
3. Indian Institute of Technology Kanpur

Int. Conference on Machine Learning (ICML) 2022

## Standard BFGS Method

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.

# Standard BFGS Method

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.
- Quasi-Newton (QN) method: $x_{t+1} = x_t - G_t^{-1} \nabla f(x_t)$.

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.
- Quasi-Newton (QN) method: $x_{t+1} = x_t - G_t^{-1} \nabla f(x_t)$.
- Standard (classical) BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

with $s_t = x_{t+1} - x_t$ and $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$.

# Standard BFGS Method

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.
- Quasi-Newton (QN) method: $x_{t+1} = x_t - G_t^{-1} \nabla f(x_t)$.
- Standard (classical) BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

with $s_t = x_{t+1} - x_t$ and $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$.

- **[A. Rodomanov and Y. Nesterov 2021 c.]** Standard BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left( \frac{d \ln \kappa}{t} \right)^{\frac{t}{2}},$$

$d$ is dimension, $\kappa$ is condition number and $\lambda_f(x)$ is Newton decrement.

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.
- Quasi-Newton (QN) method: $x_{t+1} = x_t - G_t^{-1} \nabla f(x_t)$.
- Standard (classical) BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

  with $s_t = x_{t+1} - x_t$ and $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$.

- **[A. Rodomanov and Y. Nesterov 2021 c.]** Standard BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left(\frac{d \ln \kappa}{t}\right)^{\frac{t}{2}},$$

  $d$ is dimension, $\kappa$ is condition number and $\lambda_f(x)$ is Newton decrement.

- Advantages:
  - $\Rightarrow$ Approximating the Newton direction.
  - $\Rightarrow$ Achieving superlinear convergence rate after only $d \ln \kappa$ iterations.

- Convex optimization problem: $\min_{x \in \mathbb{R}^d} f(x)$.
- Quasi-Newton (QN) method: $x_{t+1} = x_t - G_t^{-1} \nabla f(x_t)$.
- Standard (classical) BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

with $s_t = x_{t+1} - x_t$ and $y_t = \nabla f(x_{t+1}) - \nabla f(x_t)$.

- **[A. Rodomanov and Y. Nesterov 2021 c.]** Standard BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left( \frac{d \ln \kappa}{t} \right)^{\frac{t}{2}},$$

$d$ is dimension, $\kappa$ is condition number and $\lambda_f(x)$ is Newton decrement.

- Advantages:
  - ⇒ Approximating the Newton direction.
  - ⇒ Achieving superlinear convergence rate after only $d \ln \kappa$ iterations.
- Disadvantage: Failing to perfectly approximate the Hessian.

# Greedy BFGS Method

- Greedy BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t \bar{u}_t \bar{u}_t^\top G_t}{\bar{u}_t^\top G_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

# Greedy BFGS Method

- Greedy BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t \bar{u}_t \bar{u}_t^\top G_t}{\bar{u}_t^\top G_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

- $\bar{u}_t$ is the greedily selected direction:

$$\bar{u}_t = \underset{u \in \{e_i\}_{i=1}^d}{\operatorname{argmax}} \frac{u^\top G_t u}{u^\top \nabla^2 f(x_t) u},$$

where $\{e_i\}_{i=1}^d$ are the unit vectors.

- Greedy BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t \bar{u}_t \bar{u}_t^\top G_t}{\bar{u}_t^\top G_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

- $\bar{u}_t$ is the greedily selected direction:

$$\bar{u}_t = \operatorname*{argmax}_{u \in \{e_i\}_{i=1}^d} \frac{u^\top G_t u}{u^\top \nabla^2 f(x_t) u},$$

  where $\{e_i\}_{i=1}^d$ are the unit vectors.

- **[A. Rodomanov and Y. Nesterov 2021 a.]** Greedy BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left( d\kappa (1 - \frac{1}{d\kappa})^{\frac{t}{2}} \right)^t.$$

- Greedy BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t \bar{u}_t \bar{u}_t^\top G_t}{\bar{u}_t^\top G_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

- $\bar{u}_t$ is the greedily selected direction:

$$\bar{u}_t = \operatorname*{argmax}_{u \in \{e_i\}_{i=1}^d} \frac{u^\top G_t u}{u^\top \nabla^2 f(x_t) u},$$

  where $\{e_i\}_{i=1}^d$ are the unit vectors.
- **[A. Rodomanov and Y. Nesterov 2021 a.]** Greedy BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left( d\kappa (1 - \frac{1}{d\kappa})^{\frac{t}{2}} \right)^t.$$

- Advantages:
  - $\Rightarrow$ Directly approximating the Hessian matrix.
  - $\Rightarrow$ Eventually reaching fast quadratic convergence rate.

- Greedy BFGS update rule:

$$G_{t+1} = G_t - \frac{G_t \bar{u}_t \bar{u}_t^\top G_t}{\bar{u}_t^\top G_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

- $\bar{u}_t$ is the greedily selected direction:

$$\bar{u}_t = \underset{u \in \{e_i\}_{i=1}^d}{\text{argmax}} \frac{u^\top G_t u}{u^\top \nabla^2 f(x_t) u},$$

where $\{e_i\}_{i=1}^d$ are the unit vectors.

- **[A. Rodomanov and Y. Nesterov 2021 a.]** Greedy BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \le \left( d\kappa (1 - \frac{1}{d\kappa})^{\frac{t}{2}} \right)^t.$$

- Advantages:
  - $\Rightarrow$ Directly approximating the Hessian matrix.
  - $\Rightarrow$ Eventually reaching fast quadratic convergence rate.
- Disadvantage:
  - $\Rightarrow$ Requiring $d\kappa \ln(d\kappa)$ iterations to achieve the superlinear convergence.

- We proposed the sharpened BFGS method.

- We proposed the sharpened BFGS method.
- Leveraging both standard BFGS and greedy BFGS updates to
  - properly approximate the Newton direction as in BFGS.
  - accurately approximate the Hessian matrix as in Greedy BFGS.

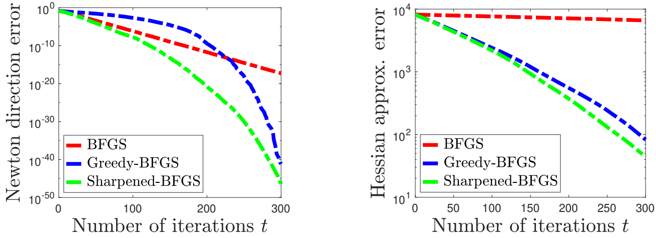
$$\bar{G}_t = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

$$G_{t+1} = \bar{G}_t - \frac{\bar{G}_t \bar{u}_t \bar{u}_t^\top \bar{G}_t}{\bar{u}_t^\top \bar{G}_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$

- We proposed the sharpened BFGS method.
- Leveraging both standard BFGS and greedy BFGS updates to
  - properly approximate the Newton direction as in BFGS.
  - accurately approximate the Hessian matrix as in Greedy BFGS.

$$\bar{G}_t = G_t - \frac{G_t s_t s_t^\top G_t}{s_t^\top G_t s_t} + \frac{y_t y_t^\top}{s_t^\top y_t},$$

$$G_{t+1} = \bar{G}_t - \frac{\bar{G}_t \bar{u}_t \bar{u}_t^\top \bar{G}_t}{\bar{u}_t^\top \bar{G}_t \bar{u}_t} + \frac{\nabla^2 f(x_t) \bar{u}_t \bar{u}_t^\top \nabla^2 f(x_t)}{\bar{u}_t^\top \nabla^2 f(x_t) \bar{u}_t}.$$



- **[Jin, Koppel, Rajawat and Mokhtari 2022]** Sharpened BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left(1 - \frac{1}{d\kappa}\right)^{\frac{t(t-1)}{4}} \left(\frac{d\kappa}{t}\right)^{\frac{t}{2}}.$$

- **[Jin, Koppel, Rajawat and Mokhtari 2022]** Sharpened BFGS method has the local superlinear convergence rate of

$$\frac{\lambda_f(x_t)}{\lambda_f(x_0)} \leq \left(1 - \frac{1}{d\kappa}\right)^{\frac{t(t-1)}{4}} \left(\frac{d\kappa}{t}\right)^{\frac{t}{2}}.$$

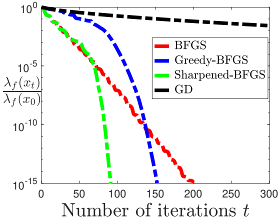
- Comparison of standard BFGS, greedy BFGS and sharpened BFGS:

| Algorithm | Superlinear Rate | $t_0$ |
|---|---|---|
| Standard BFGS | $\left(\frac{d \ln \kappa}{t}\right)^{\frac{t}{2}}$ | $d \ln \kappa$ |
| Greedy BFGS | $\left(d\kappa(1-\frac{1}{d\kappa})^{\frac{t}{2}}\right)^{t}$ | $d\kappa \ln(d\kappa)$ |
| Sharpened BFGS | $(1-\frac{1}{d\kappa})^{\frac{t(t-1)}{4}}(\frac{d\kappa}{t})^{\frac{t}{2}}$ | $d\kappa$ |

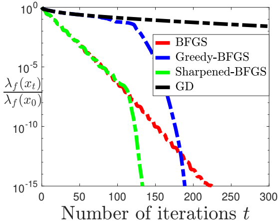
- Convergence rate of Sharpened BFGS is substantially faster than the other two methods.
- Sharpened BFGS requires less iterations compared to Greedy BFGS to enter the superlinear convergence phase.

# Numerical Experiments

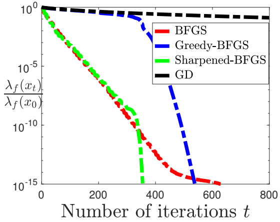
- Sharpened BFGS obtains the best performance in all considered settings.



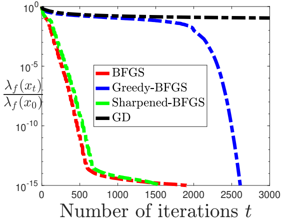

(a) phishing dataset: d = 68.




(b) a9a dataset: d = 123.




(c) protein dataset: d = 357.




(d) colon dataset: d = 2000.

Thanks for your attention!


- Q. Jin and A. Mokhtari. “**Non-asymptotic Superlinear Convergence of Standard Quasi-Newton Methods**,” *arXiv preprint arXiv:2003.13607*, 2020.
- A. Rodomanov and Y. Nesterov. “**Greedy quasi-newton methods with explicit superlinear convergence**,” *SIAM Journal on Optimization, 31(1):785–811*, 2021 a.
- A. Rodomanov and Y. Nesterov. “**Rates of superlinear con- vergence for classical quasi-newton methods**,” *Mathematical Programming, pp. 1–32*, 2021 b.
- A. Rodomanov and Y. Nesterov. “**New results on superlinear convergence of classical quasi-newton methods**,” *Journal of Optimization Theory and Applications, 188(3):744– 769*, 2021 c.
- Q.Jin, A.Koppel, K.Rajawat and A.Mokhtari . “**Sharpened Quasi-Newton Methods: Faster Superlinear Rate and Larger Local Convergence Neighborhood** ,” *The 39th International Conference on Machine Learning*, 2022.