The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

LECTURE 22

Caramanis & Mokhtari Monday, April 18, 2022

Goal: In this lecture, we study the local convergence rate of BFGS.

1 Recap

1.1 BFGS Updates

Hessian inverse approximation is given by

$$\mathbf{H}_{k+1} = (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top)\ \mathbf{H}_k\ (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top) + \rho_k \mathbf{s}_k \mathbf{s}_k^\top.$$

It is also useful to look at the sequence of Hessian approximation $\mathbf{B}_k$ update for the BFGS method, which is given by

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^\top \mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}$$

1.2 Global Convergence Analysis

We first state our assumptions:

Assumption 1. *The objective function f is twice differentiable, μ-strongly convex and its gradient is L_1-Lipschitz continuous.*

Next, we prove that the sequence $\{\cos(\theta_k)\}_{k\geq 0}$ does not converge to zero.

Lemma 1. *Consider the iterates of BFGS. There exists is a subsequence of $\{\cos(\theta_k)\}_{k\geq 0}$ that is bounded away from zero.*

Theorem 1. *Suppose we perform a descent method with the Wolfe conditions for the selection of stepsize, i.e.,*

$$f(\mathbf{x}_k + \eta_k \mathbf{p}_k) \leq f(\mathbf{x}_k) + \alpha \eta_k \nabla f(\mathbf{x}_k)^\top \mathbf{p}_k$$
$$\nabla f(\mathbf{x}_k + \eta_k \mathbf{p}_k)^\top \mathbf{p}_k \geq \beta \nabla f(\mathbf{x}_k)^\top \mathbf{p}_k$$

where $0 < \alpha < \beta < 1$. Then, if we define $\cos(\theta_k) = \frac{-\nabla f(\mathbf{x}_k)^\top \mathbf{p}_k}{\|\nabla f(\mathbf{x}_k)\| \|\mathbf{p}_k\|}$, we have

$$\sum_{k=0}^{\infty} \cos^2(\theta_k) \|\nabla f(\mathbf{x}_k)\|^2 \leq \infty.$$

By combining the above results, we obtain that for a subsequence of iterates generated by BFGS the norm of gradient approaches zero and hence they converge to the optimal solution.
The above analysis can be done more tightly to show that the following condition is satisfied.

$$\sum_{k=0}^{\infty} \|\mathbf{x}_k - \mathbf{x}^*\| < \infty.$$

2 Convergence Rates

Before stating the main results, let us first clarify what we mean by superlinear convergence rate. Suppose $\mathbf{x}_k \to \mathbf{x}^*$, and onsider the following expression

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = \rho.$$

The parameter ρ explains the speed of convergence of the sequence $\mathbf{x}_k$ to the optimal solution $\mathbf{x}^*$.

- When $\rho = 1$, we state that the convergence is **sublinear**. For instance, $\|\mathbf{x}_k - \mathbf{x}^*\| = \frac{C}{k}$ leads to a sublinear convergence rate.
- When $0 < \rho < 1$, we state that the convergence is **linear**. For instance, $\|\mathbf{x}_k - \mathbf{x}^*\| = \rho^k\|\mathbf{x}_0 - \mathbf{x}^*\|$ shows a linear convergence rate.
- When $\rho = 0$, we state that the convergence is **superlinear**. For instance, $\|\mathbf{x}_k - \mathbf{x}^*\| = (1/k)^k\|\mathbf{x}_0 - \mathbf{x}^*\|$ shows a superlinear convergence rate.

In this lecture, we show that BFGS converges at a superlinear rate.

3 Local Superlinear Convergence of BFGS

To prove this result we need a regularity condition on the objective function Hessian.

Assumption 2. *The objective function Hessian f is L_2-Lipschitz continuous, i.e.,*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\hat{\mathbf{x}})\|_2 \leq L_2\|\mathbf{x} - \hat{\mathbf{x}}\|$$

We first define a weighted version of the variable and gradient variation vectors and the Hessian approximation matrix. In particular, consider

$$\tilde{\mathbf{s}}_k := \nabla^2 f(\mathbf{x}^*)^{1/2}\mathbf{s}_k, \qquad \tilde{\mathbf{y}}_k := \nabla^2 f(\mathbf{x}^*)^{-1/2}\mathbf{y}_k, \qquad \tilde{\mathbf{B}}_k := \nabla^2 f(\mathbf{x}^*)^{-1/2}\mathbf{B}_k\nabla^2 f(\mathbf{x}^*)^{-1/2}$$

Similarly we define the modified version of the parameters that we used in the global convergence analysis. Specifically consider

$$\tilde{m}_k := \frac{\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{s}}_k}{\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{s}}_k}, \qquad \tilde{M}_k := \frac{\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{y}}_k}{\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{s}}_k}, \qquad \tilde{q}_k = \frac{\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{B}}_k\tilde{\mathbf{s}}_k}{\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{s}}_k}, \qquad \cos(\tilde{\theta}_k) = \frac{\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{B}}_k\tilde{\mathbf{s}}_k}{\|\tilde{\mathbf{s}}_k\|\|\tilde{\mathbf{B}}_k\tilde{\mathbf{s}}_k\|}.$$

Now consider the BFGS update

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top\mathbf{s}_k} - \frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k}{\mathbf{s}_k^\top\mathbf{B}_k\mathbf{s}_k}$$

If we multiply this expression from left and right by $\nabla^2 f(\mathbf{x}^*)^{-1/2}$ we obtain that

$$\tilde{\mathbf{B}}_{k+1} = \tilde{\mathbf{B}}_k + \frac{\tilde{\mathbf{y}}_k\tilde{\mathbf{y}}_k^\top}{\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{s}}_k} - \frac{\tilde{\mathbf{B}}_k\tilde{\mathbf{s}}_k\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{B}}_k}{\tilde{\mathbf{s}}_k^\top\tilde{\mathbf{B}}_k\tilde{\mathbf{s}}_k}$$

we further used the fact that $\tilde{\mathbf{y}}_k^\top\tilde{\mathbf{s}}_k = \mathbf{y}_k^\top\mathbf{s}_k$. Since the update rule for $\tilde{\mathbf{B}}_k$ and $\mathbf{B}_k$ are very similar it can be verified their Lyapunov function defined by trace-logdet are also similar and hence we have

$$\psi(\tilde{\mathbf{B}}_{k+1}) = \psi(\tilde{\mathbf{B}}_k) + (\tilde{M}_k - \ln\tilde{m}_k - 1) - \frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} + 1 + \ln\frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} + \ln\cos^2(\tilde{\theta}_k)$$

Lemma 2. *Consider the definition* $\epsilon_k := \max\{\|\mathbf{x}_k - \mathbf{x}^*\|, \|\mathbf{x}_{k+1} - \mathbf{x}^*\|\}$*. If Assumptions 1 and 2 hold, then we have*

$$\frac{\|\tilde{\mathbf{y}}_k - \tilde{\mathbf{s}}_k\|}{\|\tilde{\mathbf{s}}_k\|} \leq \tilde{c}\, \epsilon_k$$

where

$$\tilde{c} := \|\nabla^2 f(\mathbf{x}^*)^{-1}\| L_2 \leq \frac{\mu}{L_2}$$

Proof. Note that if $\mathbf{G}_k$ is the average Hessian, then we have

$$\mathbf{y}_k = \mathbf{G}_k \mathbf{s}_k$$

Hence,

$$\begin{aligned}
\tilde{\mathbf{y}}_k - \tilde{\mathbf{s}}_k &= \nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}} \mathbf{y}_k - \nabla^2 f(\mathbf{x}^*)^{\frac{1}{2}} \mathbf{s}_k \\
&= \left[\nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}} \mathbf{G}_k - \nabla^2 f(\mathbf{x}^*)^{\frac{1}{2}}\right] \mathbf{s}_k \\
&= \nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}} \left[\mathbf{G}_k - \nabla^2 f(\mathbf{x}^*)\right] \mathbf{s}_k \\
&= \nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}} \left[\mathbf{G}_k - \nabla^2 f(\mathbf{x}^*)\right] \nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}} \tilde{\mathbf{s}}_k
\end{aligned}$$

Hence, we have

$$\|\tilde{\mathbf{y}}_k - \tilde{\mathbf{s}}_k\| \leq \|\nabla^2 f(\mathbf{x}^*)^{-\frac{1}{2}}\|^2 \|\mathbf{G}_k - \nabla^2 f(\mathbf{x}^*)\| \|\tilde{\mathbf{s}}_k\|$$

Now since we can show that

$$\|\mathbf{G}_k - \nabla^2 f(\mathbf{x}^*)\| \leq L_2 \max\{\|\mathbf{x}_k - \mathbf{x}^*\|, \|\mathbf{x}_{k+1} - \mathbf{x}^*\|\}$$

the claim follows. ☐

The above result shows that as we approach the optimal solution the secant condition is almost satisfied with respect to the optimal Hessian.

Lemma 3. *For the iterates of BFGS, if* $\epsilon_k < 1/5\tilde{c}$*, then we have*

$$\tilde{m}_k := \frac{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k}{\tilde{\mathbf{s}}_k^\top \tilde{\mathbf{s}}_k} \geq 1 - \tilde{c}\epsilon_k, \qquad \tilde{M}_k := \frac{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{y}}_k}{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k} \leq 1 + 4\tilde{c}\epsilon_k$$

Proof. Note that based on the result in Lemma 1 we have

$$(1 - \tilde{c}\epsilon_k)\|\tilde{\mathbf{s}}_k\| \leq \|\tilde{\mathbf{y}}_k\| \leq (1 + \tilde{c}\epsilon_k)\|\tilde{\mathbf{s}}_k\|$$

Hence, by squaring both sides of result of Lemma 1 we have

$$\|\tilde{\mathbf{y}}_k - \tilde{\mathbf{s}}_k\|^2 \leq \tilde{c}^2\, \epsilon_k^2 \|\tilde{\mathbf{s}}_k\|^2$$

which implies that

$$\|\tilde{\mathbf{y}}_k\|^2 + \|\tilde{\mathbf{s}}_k\|^2 - 2\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k \leq \tilde{c}^2\, \epsilon_k^2 \|\tilde{\mathbf{s}}_k\|^2$$

and hence

$$\begin{aligned}
2\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k &\geq \|\tilde{\mathbf{y}}_k\|^2 + \|\tilde{\mathbf{s}}_k\|^2 - \tilde{c}^2\, \epsilon_k^2 \|\tilde{\mathbf{s}}_k\|^2 \\
&\geq (1 - \tilde{c}\epsilon_k)^2 \|\tilde{\mathbf{s}}_k\|^+ \|\tilde{\mathbf{s}}_k\|^2 - \tilde{c}^2\, \epsilon_k^2 \|\tilde{\mathbf{s}}_k\|^2 \\
&= \left[(1 - \tilde{c}\epsilon_k)^2 + 1 - \tilde{c}^2\, \epsilon_k^2\right] \|\tilde{\mathbf{s}}_k\|^2 \\
&= 2\left[1 - \tilde{c}\epsilon_k\right] \|\tilde{\mathbf{s}}_k\|^2
\end{aligned}$$

Hence, the first claim easily follows and we have

$$\tilde{m}_k := \frac{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k}{\tilde{\mathbf{s}}_k^\top \tilde{\mathbf{s}}_k} \geq 1 - \tilde{c}\epsilon_k$$

Now note that we have

$$\tilde{M}_k := \frac{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{y}}_k}{\tilde{\mathbf{y}}_k^\top \tilde{\mathbf{s}}_k} \leq \frac{(1+\tilde{c}\epsilon_k)^2 \|\tilde{\mathbf{s}}_k\|^2}{(1-\tilde{c}\epsilon_k)\|\tilde{\mathbf{s}}_k\|^2} = \frac{(1+\tilde{c}\epsilon_k)^2}{(1-\tilde{c}\epsilon_k)}$$

Now since we have $\epsilon_k \tilde{c} < 1/5$ we have

$$\frac{(1+\tilde{c}\epsilon_k)^2}{(1-\tilde{c}\epsilon_k)} \leq 1 + 4\tilde{c}\epsilon_k \iff 1 + 2\tilde{c}\epsilon_k + \tilde{c}^2\epsilon_k^2 \leq 1 + 3\tilde{c}\epsilon_k - 4\tilde{c}^2\epsilon_k^2 \iff 5\tilde{c}^2\epsilon_k^2 \leq \tilde{c}\epsilon_k$$

and the second claim also follows. □

Lemma 4. *If we have $\epsilon_k \tilde{c} < 1/2$, then*

$$\lim_{k\to\infty} \cos(\tilde{\theta}_k) = 1, \qquad \lim_{k\to\infty} \tilde{q}_k = 1.$$

Proof. We can further show that if $\epsilon_k \tilde{c} < 1/2$ then we have

$$\ln(1 - \epsilon_k \tilde{c}) \geq -2\epsilon_k \tilde{c}$$

which implies that

$$\ln(m_k) \geq -2\epsilon_k \tilde{c}$$

Hence, using the results in Lemma 1 and Lemma 3 we have

$$\psi(\tilde{\mathbf{B}}_{k+1}) \leq \psi(\tilde{\mathbf{B}}_k) + 5\epsilon_k \tilde{c} + \ln\cos^2(\tilde{\theta}_k) + \left[1 - \frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} + \ln\frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)}\right]$$

By summing up the terms and using the fact that $\sum_{j=0}^{\infty} \|\mathbf{x}_j - \mathbf{x}^*\| < \infty$ we obtain that

$$\sum_{k=0}^{\infty} \ln\frac{1}{\cos^2(\tilde{\theta}_k)} - \left[1 - \frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} + \ln\frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)}\right] \leq \psi(\tilde{\mathbf{B}}_0) + 5\tilde{c}\sum_{k=0}^{\infty} \epsilon_k < \infty$$

Hence,

$$\lim_{k\to\infty} \frac{1}{\cos^2(\tilde{\theta}_k)} = 0, \qquad \lim_{k\to\infty} 1 - \frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} + \ln\frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} = 0$$

which lead to

$$\lim_{k\to\infty} \cos(\tilde{\theta}_k) = 1, \qquad \lim_{k\to\infty} \tilde{q}_k = 1.$$

□

Finally, we state the main throem.

Theorem 2. *Suppose* $\epsilon_k < \frac{L_2}{5\mu}$, *where* $\epsilon_k := \max\{\|\mathbf{x}_k - \mathbf{x}^*\|, \|\mathbf{x}_{k+1} - \mathbf{x}^*\|\}$. *If Assumptions 1 and 2 hold, then we have*

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$

Proof.

$$\begin{aligned}
\frac{\|\nabla^2 f(\mathbf{x}^*)^{-1/2}(\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*))\mathbf{s}_k\|^2}{\|\nabla^2 f(\mathbf{x}^*)^{-1/2}\mathbf{s}_k\|^2} &= \frac{\|(\tilde{\mathbf{B}}_k - \mathbf{I})\tilde{\mathbf{s}}_k\|^2}{\|\tilde{\mathbf{s}}_k\|^2} \\
&= \frac{\|\tilde{\mathbf{B}}\tilde{\mathbf{s}}_k\|^2 - 2\tilde{\mathbf{s}}_k^\top \tilde{\mathbf{B}}_k \tilde{\mathbf{s}}_k + \|\tilde{\mathbf{s}}_k\|^2}{\|\tilde{\mathbf{s}}_k\|^2} \\
&= \frac{\tilde{q}_k}{\cos^2(\tilde{\theta}_k)} - 2\tilde{q}_k + 1
\end{aligned}$$

Hence,

$$\lim_{k\to\infty} \frac{\|\nabla^2 f(\mathbf{x}^*)^{-1/2}(\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*))\mathbf{s}_k\|^2}{\|\nabla^2 f(\mathbf{x}^*)^{-1/2}\mathbf{s}_k\|^2} = 0 \;\Rightarrow\; \lim_{k\to\infty} \frac{\|(\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*))\mathbf{s}_k\|^2}{\|\mathbf{s}_k\|^2} = 0$$

Now note that

$$\begin{aligned}
(\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*))\mathbf{s}_k &= -\nabla f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}^*)(\mathbf{x}_{k+1} - \mathbf{x}_k) \\
&= \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}^*)(\mathbf{x}_{k+1} - \mathbf{x}_k) - \nabla f(\mathbf{x}_{k+1})
\end{aligned}$$

Which implies that

$$\|\nabla f(\mathbf{x}_{k+1})\| \le \|(\mathbf{B}_k - \nabla^2 f(\mathbf{x}^*))\mathbf{s}_k\| + \|\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) - \nabla^2 f(\mathbf{x}^*)(\mathbf{x}_{k+1} - \mathbf{x}_k)\|$$

Hence, we have

$$\lim_{k\to\infty} \frac{\|\nabla f(\mathbf{x}_{k+1})\|}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} = 0$$

Now since we have

$$\frac{\|\nabla f(\mathbf{x}_{k+1})\|}{\|\mathbf{x}_{k+1} - \mathbf{x}_k\|} \ge \frac{\|\nabla f(\mathbf{x}_{k+1})\|}{\|\mathbf{x}_{k+1} - \mathbf{x}^*\| + \|\mathbf{x}_k - \mathbf{x}^*\|} \ge \frac{\mu\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_{k+1} - \mathbf{x}^*\| + \|\mathbf{x}_k - \mathbf{x}^*\|} = \mu \frac{\frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|}}{1 + \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|}}$$

Hence, we have

$$\lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|} = 0$$

$\square$