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

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

Lecture 22

Caramanis & Mokhtari Wednesday, April 13, 2022

Goal: In this lecture, we talk about the convergence properties of the BFGS method. For more detailed analysis check

A TOOL FOR THE ANALYSIS OF QUASI-NEWTON METHODS WITH APPLICATION TO UNCONSTRAINED MINIMIZATION

1 BFGS

If we consider $\mathbf{H}_k = \mathbf{B}_k^{-1}$ in the BFGS update we have

$$\begin{aligned} &\text{minimize} && \|\mathbf{H}-\mathbf{H}_k\|_{\mathbf{G}_k}^2 \\ &\text{subject to} && \mathbf{H}^\top = \mathbf{H} \qquad \mathbf{s}_k = \mathbf{H}\mathbf{y}_k \end{aligned}$$

the unique solution of the problem would be

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

which provides an update rule for the sequence of Hessian inverse approximation matrices!

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

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

We further showed that

$$\begin{aligned} \mathbf{B}_{k+1}^{BFGS} = &\operatorname*{argmin}_{\mathbf{X}} \operatorname{tr}(\mathbf{B}_k^{-1}\mathbf{X}) - \ln\det(\mathbf{B}_k^{-1}\mathbf{X}) - n \\ &\text{subject to} \quad \mathbf{X}\mathbf{s}_k = \mathbf{y}_k, \qquad \mathbf{X} \succeq \mathbf{0}. \end{aligned}$$

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.*

Lemma 1. *If Assumption 1 holds, then*

$$\frac{\mathbf{y}_k^\top\mathbf{s}_k}{\mathbf{s}_k^\top\mathbf{s}_k} \geq \mu, \qquad \frac{\mathbf{y}_k^\top\mathbf{y}_k}{\mathbf{y}_k^\top\mathbf{s}_k} \leq L_1,$$

Proof. Note that if $\hat{\mathbf{G}}_k$ is the average Hessian between $\mathbf{x}_k$ and $\mathbf{x}_{k+1}$ then we have $\mathbf{y}_k = \hat{\mathbf{G}}_k\mathbf{s}_k$. Hence,

$$\frac{\mathbf{y}_k^\top \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{s}_k} = \frac{\mathbf{s}_k^\top \hat{\mathbf{G}}_k \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{s}_k} \geq \mu$$

and

$$\frac{\mathbf{y}_k^\top \mathbf{y}_k}{\mathbf{y}_k^\top \mathbf{s}_k} = \frac{\mathbf{s}_k^\top \hat{\mathbf{G}}_k \hat{\mathbf{G}}_k \mathbf{s}_k}{\mathbf{s}_k^\top \hat{\mathbf{G}}_k \mathbf{s}_k} = \frac{\mathbf{z}_k^\top \hat{\mathbf{G}}_k \mathbf{z}_k}{\mathbf{z}_k^\top \mathbf{z}_k} \leq L_1$$

where we used $\mathbf{z}_k = \hat{\mathbf{G}}_k^{1/2}\mathbf{s}_k$ □

This result implies that if we define

$$m_k := \frac{\mathbf{y}_k^\top \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{s}_k} \qquad M_k := \frac{\mathbf{y}_k^\top \mathbf{y}_k}{\mathbf{y}_k^\top \mathbf{s}_k}$$

then we have

$$m_k \geq \mu, \qquad M_k \leq L_1$$

Lemma 2. *For the BFGS update we have*

$$\text{trace}(\mathbf{B}_{k+1}) = \text{trace}(\mathbf{B}_k) + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\|\mathbf{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}, \qquad \det(\mathbf{B}_{k+1}) = \det(\mathbf{B}_k)\frac{\mathbf{y}_k^\top \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}.$$

Proof.

$$\begin{aligned}\text{trace}(\mathbf{B}_{k+1}) &= \text{trace}(\mathbf{B}_k) + \text{trace}(\frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k}) - \text{trace}(\frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}) \\ &= \text{trace}(\mathbf{B}_k) + \frac{\|\mathbf{y}_k\|^2}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\|\mathbf{B}_k\mathbf{s}_k\|^2}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\end{aligned}$$

Now using the fact that

$$\det(\mathbf{I} + \mathbf{a}\mathbf{b}^\top + \mathbf{u}\mathbf{v}^\top) = (1 + \mathbf{a}^\top\mathbf{b})(1 + \mathbf{u}^\top\mathbf{v}) - (\mathbf{a}^\top\mathbf{u})(\mathbf{b}^\top\mathbf{v})$$

we can show that

$$\begin{aligned}\det(\mathbf{B}_{k+1}) &= \det\left(\mathbf{B}_k + \frac{\mathbf{y}_k\mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right) \\ &= \det\left(\mathbf{B}_k^{1/2}\left(\mathbf{I} + \frac{\mathbf{B}_k^{-1/2}\mathbf{y}_k\mathbf{y}_k^\top\mathbf{B}_k^{-1/2}}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k^{1/2}\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k^{1/2}}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right)\mathbf{B}_k^{1/2}\right) \\ &= \det(\mathbf{B}_k^{1/2})^2 \det\left(\mathbf{I} + \frac{\mathbf{B}_k^{-1/2}\mathbf{y}_k\mathbf{y}_k^\top\mathbf{B}_k^{-1/2}}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k^{1/2}\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k^{1/2}}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right) \\ &= \det(\mathbf{B}_k) \det\left(\mathbf{I} + \frac{\mathbf{B}_k^{-1/2}\mathbf{y}_k\mathbf{y}_k^\top\mathbf{B}_k^{-1/2}}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k^{1/2}\mathbf{s}_k\mathbf{s}_k^\top\mathbf{B}_k^{1/2}}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right) \\ &= \det(\mathbf{B}_k)\left[\left(1 + \frac{\mathbf{y}_k^\top\mathbf{B}_k^{-1}\mathbf{y}_k}{\mathbf{y}_k^\top \mathbf{s}_k}\right)\left(1 - \frac{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right) + \frac{(\mathbf{y}_k^\top\mathbf{B}_k^{-1/2}\mathbf{B}_k^{1/2}\mathbf{s}_k)^2}{\mathbf{y}_k^\top \mathbf{s}_k\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\right] \\ &= \det(\mathbf{B}_k)\frac{\mathbf{y}_k^\top \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}\end{aligned}$$

□

Now define the new parameters

$$q_k = \frac{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}{\mathbf{s}_k^\top \mathbf{s}_k}, \qquad \cos(\theta_k) = \frac{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}{\|\mathbf{s}_k\| \|\mathbf{B}_k \mathbf{s}_k\|}$$

Lemma 3. *For the BFGS updates we have*

$$\psi(\mathbf{B}_{k+1}) \leq \psi(\mathbf{B}_k) + (L_1 - \ln \mu - 1) + \ln \cos^2(\theta_k)$$

Proof. Using these definitions we have

$$\text{trace}(\mathbf{B}_{k+1}) = \text{trace}(\mathbf{B}_k) + M_k - \frac{q_k}{\cos^2(\theta_k)}, \qquad \det(\mathbf{B}_{k+1}) = \det(\mathbf{B}_k) \frac{m_k}{q_k}.$$

Hence, if we define

$$\psi(\mathbf{B}) = \text{trace}(\mathbf{B}) - \ln \det(\mathbf{B})$$

then we have

$$\psi(\mathbf{B}_{k+1}) = \psi(\mathbf{B}_k) + M_k - \frac{q_k}{\cos^2(\theta_k)} - \ln m_k + \ln q_k$$

which can be written as

$$\begin{aligned} \psi(\mathbf{B}_{k+1}) &= \psi(\mathbf{B}_k) + (M_k - \ln m_k - 1) - \frac{q_k}{\cos^2(\theta_k)} + 1 + \ln q_k + \ln \frac{q_k}{\cos^2(\theta_k)} - \ln \frac{q_k}{\cos^2(\theta_k)} \\ &= \psi(\mathbf{B}_k) + (M_k - \ln m_k - 1) - \frac{q_k}{\cos^2(\theta_k)} + 1 + \ln \frac{q_k}{\cos^2(\theta_k)} + \ln \cos^2(\theta_k) \end{aligned}$$

Since $1 - t + \ln t \leq 0$ we have

$$\begin{aligned} \psi(\mathbf{B}_{k+1}) &\leq \psi(\mathbf{B}_k) + (M_k - \ln m_k - 1) + \ln \cos^2(\theta_k) \\ &\leq \psi(\mathbf{B}_k) + (L_1 - \ln \mu - 1) + \ln \cos^2(\theta_k). \end{aligned}$$

□

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

Lemma 4. *Consider the iterates of BFGS. There exists is a subsequence of $\{\cos(\theta_k)\}_{k \geq 0}$ that is bounded away from zero.*

Proof. Let us define $C_1 = L_1 - \ln \mu - 1$. Then, using the above expression we have

$$0 < \psi(\mathbf{B}_{k+1}) \leq \psi(\mathbf{B}_0) + C_1(k+1) + \sum_{j=0}^{k} \ln \cos^2(\theta_j)$$

Now we prove by contradiction that $\lim_{k \to \infty} \cos^2(\theta_k) \neq 0$. Suppose $\lim_{k \to \infty} \cos^2(\theta_k) = 0$. Then, there exists k_1 such that for $j > k_1$ we have

$$\ln \cos^2(\theta_j) < -2C_1$$

Hence, we have

$$0 \leq \psi(\mathbf{B}_0) + C_1(k+1) - 2C_1(k - k_1) + \sum_{j=0}^{k_1} \ln \cos^2(\theta_j)$$

which implies

$$0 \leq \psi(\mathbf{B}_0) + 2C_1k_1 - C_1(k-1) + \sum_{j=0}^{k_1} \ln\cos^2(\theta_j)$$

Now if we send k to ∞ then the right hand side becomes zero and hence we reach contradiction.

This observation shows that for any $\delta > 0$, there exists a subsequence of iterates where their indices is denoted by J_k such that $\ln\cos^2(\theta_j) > -2C_1$ which implies that $\cos(\theta_{j_k}) > e^{-C_1}$. □

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*

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

Proof. From the second condition we have

$$(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k))^\top\mathbf{p}_k \geq (\beta-1)\nabla f(\mathbf{x}_k)^\top\mathbf{p}_k$$

From the Lipschitz continuity of the gradient we have

$$(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k))^\top\mathbf{p}_k \leq \|(\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k))\|\|\mathbf{p}_k\| \leq \eta_k L_1\|\mathbf{p}_k\|^2$$

Hence,

$$\eta_k \geq \frac{(\beta-1)\nabla f(\mathbf{x}_k)^\top\mathbf{p}_k}{L_1\|\mathbf{p}_k\|^2}$$

Now using the first condition we have

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha\frac{(1-\beta)(\nabla f(\mathbf{x}_k)^\top\mathbf{p}_k)^2}{L_1\|\mathbf{p}_k\|^2}$$

which can be written as

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha\frac{(1-\beta)}{L_1}\cos^2(\theta_k)\|\nabla f(\mathbf{x}_k)\|^2$$

and the claim easily follows. □

By combining the results in Lemma 4 and Theorem 1, 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.

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