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

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

Lecture 21

Caramanis & Mokhtari Monday, April 11, 2022

Goal: In this lecture, we talk about quasi-Newton methods.

1 Main Idea

As we discussed in previous lecture, the update of Newton's method is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$$

Implementation of this algorithm requires computing the Hessian and solving a linear system of the form $\nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) = -\eta_k \nabla f(\mathbf{x}_k)$. Both of these computations could be very costly, when the dimension of the problem n is large. Specifically in comparison to gradient-based methods.

The main idea of quasi-Newton methods is to approximate the Newton direction $\nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)$ without computing the Hessian or its inverse and simply by using first-oder information, i.e., gradients.

In its most general form, the update of quasi-Newton methods is given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{B}_k^{-1} \nabla f(\mathbf{x}_k)$$

where $\mathbf{B}_k$ is a matrix that approximates the Hessian $\nabla^2 f(\mathbf{x}_k)$. There are several different schemes for selection of the Hessian approximation matrix $\mathbf{B}_k$ and each leads to a different quasi-Newton method.

2 Secant Condition

Note that an easy way to approximate second derivative for a scalar function is

$$f''(\mathbf{x}_k) \approx \frac{f'(\mathbf{x}_k) - f'(\mathbf{x}_{k-1})}{\mathbf{x}_k - \mathbf{x}_{k-1}}$$

The same idea can be extended to higher-dimension by saying that

$$\nabla^2 f(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}_{k-1}) \approx \nabla f(\mathbf{x}_k) - \nabla f(\mathbf{x}_{k-1})$$

Hence, we would like our Hessian approximation to satisfy the following condition

$$\mathbf{B}_k(\mathbf{x}_k - \mathbf{x}_{k-1}) = \mathbf{x}_k - \mathbf{x}_{k-1}$$

This way at least approximately the Hessian approximation $\mathbf{B}_k$ is in agreement with the true Hessian $\nabla^2 f(\mathbf{x}_k)$ towards the direction $\mathbf{x}_k - \mathbf{x}_{k-1}$.

3 Broyden's method

In Broyden's method the main idea is to keep two consecutive Hessian approximations close, while the new one satisfies the secant condition.

$$\begin{array}{ll} \text{minimize} & \|\mathbf{B} - \mathbf{B}_k\|_F^2 \\ \text{subject to} & \mathbf{B}(\mathbf{x}_{k+1} - \mathbf{x}_k) = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k) \end{array}$$

Before stating the solution of the problem, let us define the variable and gradient variations as

$$\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k, \qquad \mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$$

Considering these definitions, one can show that the solution of the above optimization problem is

$$\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{1}{\|\mathbf{s}_k\|^2}(\mathbf{y}_k - \mathbf{B}_k\mathbf{s}_k)\mathbf{s}_k^\top$$

(why? Exercise!)
The above update changes $\mathbf{B}_k$ using a rank-one update and it can be easily verified that it leads to a matrix $\mathbf{B}_{k+1}$ that satisfies the secant condition, i.e., $\mathbf{B}_{k+1}\mathbf{y}_k = \mathbf{s}_k$. Now using the Sherman-Morrison formula one can establish an update for the Hessian inverse approximation matrices as

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

A major issue with the update of Broyden's method is that it is not symmetric and hence the sequence of Hessian approximation matrices may not stay symmetric and positive definite.

Hence, Broyden's method can be summarized as follows:

Initialization Step. Select $\mathbf{B}_0$ and $\mathbf{x}_0$

Main Loop.

- for $k = 0, 1, \ldots$
  - Compute the objective function gradient $\nabla f(\mathbf{x}_k)$
  - Compute the new iterate $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k\mathbf{B}_k^{-1}\nabla f(\mathbf{x}_k)$
  - Compute $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
  - Compute the new Hessian inverse approximation $\mathbf{B}_{k+1}^{-1} = \mathbf{B}_k^{-1} + \frac{1}{\mathbf{s}_k^\top\mathbf{B}_k^{-1}\mathbf{y}_k}(\mathbf{s}_k - \mathbf{B}_k^{-1}\mathbf{y}_k)\mathbf{s}_k^\top\mathbf{B}_k^{-1}$

4 DFP

To keep the Hessian approximation matrices symmetric we need to enforce this condition into the optimization problem that we solve for the new Hessian approximation. This can be done by solving the following problem

$$\begin{array}{ll} \text{minimize} & \|\mathbf{B} - \mathbf{B}_k\|_{\mathbf{W}}^2 \\ \text{subject to} & \mathbf{B}^\top = \mathbf{B} \qquad \mathbf{B}\mathbf{s}_k = \mathbf{y}_k \end{array}$$

In the above expression we replaced the Frobenius norm by some weighted matrix norm to keep the subproblem as general as possible, where

$$\|\mathbf{A}\|_{\mathbf{W}} = \|\mathbf{W}^{1/2}\mathbf{A}\mathbf{W}^{1/2}\|_{\mathbf{F}}$$

Note that the only condition that we require for $\mathbf{W}$ is to satisfy the condition $\mathbf{y}_k = \mathbf{W}\mathbf{s}_k$. This way we end up with a scale-invariant solution. Specifically, we can choose the weight matrix as $\mathbf{W} = \mathbf{G}_k^{-1}$ where

$$\mathbf{G}_k = \int_0^1 \nabla^2 f(\tau\mathbf{x}_k + (1-\tau)\mathbf{x}_{k+1}) \quad d\tau$$

is the average Hessian matrix and by Taylor's theorem satisfies the secant condition, i.e., $\mathbf{y}_k = \mathbf{G}_k\mathbf{s}_k$. Considering this discussion a valid subproblem for Hessian approximation is

$$\begin{aligned} &\text{minimize} && \|\mathbf{B} - \mathbf{B}_k\|^2_{\mathbf{G}_k^{-1}} \\ &\text{subject to} && \mathbf{B}^\top = \mathbf{B} \qquad \mathbf{B}\mathbf{s}_k = \mathbf{y}_k \end{aligned}$$

The unique solution of the above problem can be written as

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

where

$$\rho_k = \frac{1}{\mathbf{y}_k^\top\mathbf{s}_k}$$

This algorithm was first proposed by Davidon and then was analyzed by Fletcher and Powell and for these reasons it is often called the DFP method.
Note that using Sherman-Morrison formula we again can write an explicit update for the Hessian inverse approximation matrices as

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

Note that the last two terms are rank one matrices. Hence, the update of DFP requires a rank two update that keeps the matrices symmetric while satisfying the secant condition!

Note that the update of DFP doesn't require any matrix inversion as we directly update the Hessian inverse approximation matrices. Next, we summarize the steps of the DFP method:

Initialization Step. Select $\mathbf{B}_0^{-1}$ and $\mathbf{x}_0$

Main Loop.

- for $k = 0, 1, \dots$
  - Compute the objective function gradient $\nabla f(\mathbf{x}_k)$
  - Compute the new iterate $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k\mathbf{B}_k^{-1}\nabla f(\mathbf{x}_k)$
  - Compute $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
  - Compute the new Hessian inverse approximation

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

5 BFGS

Next, we present the dual of the DFP method. The main idea of the BFGS method is very similar to the DFP method, except the fact that in BFGS we minimize the distance between two consecutive Hessian inverse approximation matrices, instead of the distance between the Hessian approximation matrices in DFP. To be more precise, if we consider $\mathbf{H}_k = \mathbf{B}_k^{-1}$ in the BFGS update we have

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

where in this case $\mathbf{W}$ should be such $\mathbf{W}\mathbf{s}_k = \mathbf{y}_k$. In this case, if we simply choose the weight matrix as $\mathbf{W} = \mathbf{G}_k$ which leads to

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

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

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

This update is not useful for practical reasons, but it comes handy in the analysis of the BFGS method.
This algorithm was called BFGS for its discoveries by Broyden, Fletcher, Goldfarb, and Shanno.

Note that the update of BFGS also doesn't require any matrix inversion as we directly update the Hessian inverse approximation matrices. Next, we summarize the steps of the BFGS method:

Initialization Step. Select $\mathbf{H}_0$ and $\mathbf{x}_0$

Main Loop.

- for $k = 0, 1, \ldots$
  - Compute the objective function gradient $\nabla f(\mathbf{x}_k)$
  - Compute the new iterate $\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \mathbf{H}_k \nabla f(\mathbf{x}_k)$
  - Compute $\mathbf{s}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$
  - Compute the new Hessian inverse approximation

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

6 Important properties of BFGS

Note that the updates of BFGS and DFP are properly defined if $\mathbf{y}_k^\top \mathbf{s}_k$ is bounded away from zero. In the following lemma, we show that this condition holds if f is strictly convex.

Lemma 1. *Recall $\mathbf{s}_k$ and $\mathbf{y}_k$. Suppose $\mathbf{s}_k$ is nonzero. If f is strictly convex then we have*

$$\mathbf{y}_k^\top \mathbf{s}_k > 0$$

and if f is μ-strongly convex then we have

$$\mathbf{y}_k^\top \mathbf{s}_k \geq \mu \|\mathbf{s}_k\|^2$$

Proof. The proof simply follows from the definitions of strict convexity and strong convexity. □

Note that $\mathbf{s}_k$ is zero if and only if $\nabla f(\mathbf{x}_k) = \mathbf{0}$ which holds if and only if $\mathbf{x}_k = \mathbf{x}^*$, for strongly convex functions.

Lemma 2. *If the initial Hessian approximation $\mathbf{H}_0$ is symmetric positive definite and f is strongly convex, then all matrices $\mathbf{H}_k$ are symmetric positive definite.*

Proof. Note that for any unit vector $\mathbf{v}$ we have

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

Now both terms are indeed non-negative. Further not that the first term is only zero if $\mathbf{v}$ is parallel to $\mathbf{y}_k$, and in this case the second term is indeed positive.

□

7 Optimality of BFGS and DFP

A common approach for measuring distance between two positive definite matrices is by looking at their Gaussian differential entropy. Consider two Gaussian distributions of the form $N(\mathbf{0}, \mathbf{X})$ and $N(\mathbf{0}, \mathbf{Y})$. Their differential entropy is defined as

$$\operatorname{tr}(\mathbf{X}\mathbf{Y}^{-1}) - \log\det(\mathbf{X}\mathbf{Y}^{-1}) - n$$

It can be easily verified that this function is positive for any PD matrix $\mathbf{X}$ and it is zero when $\mathbf{X} = \mathbf{Y}$. It can be shown that BFGS minimizers the differential entropy between $N(\mathbf{0}, \mathbf{X})$ and $N(\mathbf{0}, \mathbf{B}_k)$ while satisfying the secant condition. In another words,

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

Exercise: Show that $\mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k \mathbf{y}_k^\top}{\mathbf{y}_k^\top \mathbf{s}_k} - \frac{\mathbf{B}_k \mathbf{s}_k \mathbf{s}_k^\top \mathbf{B}_k}{\mathbf{s}_k^\top \mathbf{B}_k \mathbf{s}_k}$ is the unique solution of the above problem.

A similar argument holds for the DFP method. Specifically, for DFP, we aim to minimize the differential entropy between $N(\mathbf{0}, \mathbf{B}_k)$ and $N(\mathbf{0}, \mathbf{X})$ while satisfying the secant condition, i.e.,

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

Exercise: Show that $\mathbf{B}_{k+1} = (\mathbf{I} - \rho_k \mathbf{y}_k \mathbf{s}_k^\top) \, \mathbf{B}_k \, (\mathbf{I} - \rho_k \mathbf{s}_k \mathbf{y}_k^\top) + \rho_k \mathbf{y}_k \mathbf{y}_k^\top$ is the unique solution of the above problem.