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

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

Lecture 18

Caramanis & Mokhtari — Wednesday, March 30, 2022

Goal: In this lecture we talk about self-concordant functions and the convergence rate of Newton's method for this class of functions.

1 Newton is affine-invariant

One important observation that once has to have in mind is that Newton's method is affine invariant. Consider $f(\mathbf{x})$ and $\phi(\mathbf{y}) = f(\mathbf{B}\mathbf{y})$.
Further, consider the updates

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

and

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \nabla^2 \phi(\mathbf{y}_k)^{-1} \nabla \phi(\mathbf{y}_k)$$

It can be shown that if $\mathbf{x}_0 = \mathbf{B}\mathbf{y}_0$, then for all $k \geq 0$ we have $\mathbf{x}_k = \mathbf{B}\mathbf{y}_k$

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

Hence, the Newton method is affine-invariant. However, our previous analysis does not have this property. The main issue is the following assumption on the Hessian is not affine-invariant:

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

Hence, in this lecture we address this issue.

2 Self-concordant functions

To get rid of the above issue we will focus on affine-invariant functions.

Definition 1. *A convex function $f : \mathbb{R} \to \mathbb{R}$ is called self-concordant if there exists a constant $M_f \geq 0$ such that*

$$|f'''(x)| \leq 2M_f f''(x)^{3/2}$$

If $M_f = 1$, then the function is called standard self-concordant.

This definition can be extended to $\mathbb{R}^n$.

Definition 2. *A convex function $f : \mathbb{R}^n \to \mathbb{R}$ is called self-concordant if it is self-concordant along every direction line in its domain. In other words, f is self-concordant if $\phi(t) := f(\mathbf{x} + t\mathbf{u})$ is a self-concordant function for all $\mathbf{x}$ and $\mathbf{u}$.This is equivalent to*

$$|D^3 f(\mathbf{x})[\mathbf{u},\mathbf{u},\mathbf{u}]| \leq 2M_f(D^2 f(\mathbf{x})[\mathbf{u},\mathbf{u}])^{3/2} = 2M_f(\mathbf{u}^\top \nabla^2 f(\mathbf{x})\mathbf{u})^{3/2}$$

Why this definition is important? Because it is invariant to changes in coordinations. In other words, if $f(x)$ is M_f self-concordant, then $\phi(y) = f(ay+b)$ is also M_f self-concordant.

$$|\phi(y)'''| = |a^3 f'''(ay+b)| \leq 2M_f(a^2 f''(ay+b))^{3/2} = 2M_f(\phi(y)'')^{3/2}$$

2.1 Examples of self-concordant functions

1. linear functions: $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + c$
2. quadratic functions $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$
3. logarithmic functions of the form $f(x) = -\log(x)$

2.2 Properties of self-concordant functions

Note that this condition provides an upper bound on the third derivative of the functions. Hence, it controls the variation of the Hessian by an upper bound that also depends on the Hessian. This observation leads to the following result which provides an upper bound on the variation of the Hessian for self-concordant functions.

Lemma 1. *If $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant, then we have*

$$(1 - M_f r)^2\ \nabla^2 f(\mathbf{x})\ \preceq\ \nabla^2 f(\mathbf{y})\ \preceq\ \frac{1}{(1-M_f r)^2}\ \nabla^2 f(\mathbf{x})$$

where $r = \|\mathbf{y}-\mathbf{x}\|_{\nabla^2 f(\mathbf{x})} = \sqrt{(\mathbf{y}-\mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y}-\mathbf{x})}$.

Proof. Exercise. □

Lemma 2. *If $f : \mathbb{R}^n \to \mathbb{R}$ is self-concordant and $r = \|\mathbf{y}-\mathbf{x}\|_{\nabla^2 f(\mathbf{x})} \leq 1/M_f$ then we have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) + \frac{1}{M_f^2}\omega_*(M_f\|\mathbf{y}-\mathbf{x}\|_{\nabla^2 f(\mathbf{x})})$$

where $\omega_*(t) = -t - ln(1-t)$

Proof. Exercise. □

Hence, we have

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y}-\mathbf{x}) - \frac{1}{M_f}\|\mathbf{y}-\mathbf{x}\|_{\nabla^2 f(\mathbf{x})} - \frac{1}{M_f^2}\ln(1 - M_f\|\mathbf{y}-\mathbf{x}\|_{\nabla^2 f(\mathbf{x})})$$

3 Convergence analysis of Newton's method

Recall the update of Newton's method with the backtracking line-search where we have

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

and recall the definition of Newton decrement defined as

$$\lambda(\mathbf{x}_k) = \sqrt{\nabla f(\mathbf{x}_k)^\top \nabla^2 f(\mathbf{x}_k)^{-1} \nabla f(\mathbf{x}_k)}$$

Lemma 3. *In the update of Newton's method we have*

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

and for the case that $\eta_k \lambda(\mathbf{x}_k) \leq 1/M_f$ *we have*

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \eta_k \lambda(\mathbf{x}_k)^2 - \frac{\eta_k \lambda(\mathbf{x}_k)}{M_f} - \frac{1}{M_f^2} \ln(1 - M_f \eta_k \lambda(\mathbf{x}_k))$$

Proof. It can be easily verified that, for the Newton's method we have

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

Hence, we have

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

Now using the result in Lemma 2 we have if $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{\nabla^2 f(\mathbf{x}_k)} = \eta_k \lambda(\mathbf{x}_k) \leq 1/M_f$ then we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top (\mathbf{x}_{k+1} - \mathbf{x}_k) - \frac{1}{M_f} \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{\nabla^2 f(\mathbf{x}_k)} - \frac{1}{M_f^2} \ln(1 - M_f \|\mathbf{x}_{k+1} - \mathbf{x}_k\|_{\nabla^2 f(\mathbf{x}_k)}) \\ &= f(\mathbf{x}_k) - \eta_k \lambda(\mathbf{x}_k)^2 - \frac{\eta_k \lambda(\mathbf{x}_k)}{M_f} - \frac{1}{M_f^2} \ln(1 - M_f \eta_k \lambda(\mathbf{x}_k)) \end{aligned}$$

□

Next, we study the analysis of Newton's method. We divide the analysis into two cases:

- Global Phase: $M_f \lambda(\mathbf{x}_k) \geq (1 - 2\alpha)/4$. We show that in this case, we have

  $$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_{k+1}) - \gamma$$

  where $\gamma = \frac{\alpha\beta((1-2\alpha)/4M_f)^2}{1+M_f((1-2\alpha)/4M_f)}$

- Local phase: $M_f \lambda(\mathbf{x}_k) < (1 - 2\alpha)/4$. In this phase we have a quadratic convergence rate.

4 Global Analysis

We first show that $\hat{\eta}_k = \frac{1}{1+M_f\lambda(\mathbf{x}_k)}$ is always admissible in the global phase. First, note that $\hat{\eta}_k$ satisfies the required condition in the above lemma, hence we can show that if $\mathbf{x}_{k+1} = \mathbf{x}_k + \hat{\eta}_k\Delta_{newton}$, then we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \frac{\lambda(\mathbf{x}_k)^2}{1+M_f\lambda(\mathbf{x}_k)} - \frac{\lambda(\mathbf{x}_k)}{M_f(1+M_f\lambda(\mathbf{x}_k))} - \frac{1}{M_f^2}\ln(1-\frac{M_f\lambda(\mathbf{x}_k)}{1+M_f\lambda(\mathbf{x}_k)}) \\
&= f(\mathbf{x}_k) - \frac{\lambda(\mathbf{x}_k)^2}{1+M_f\lambda(\mathbf{x}_k)} - \frac{M_f\lambda(\mathbf{x}_k)}{M_f^2(1+M_f\lambda(\mathbf{x}_k))} - \frac{1}{M_f^2}\ln\left(1-\frac{M_f\lambda(\mathbf{x}_k)}{1+M_f\lambda(\mathbf{x}_k)}\right) \\
&= f(\mathbf{x}_k) - \frac{\lambda(\mathbf{x}_k)}{M_f}\frac{M_f\lambda(\mathbf{x}_k)}{1+M_f\lambda(\mathbf{x}_k)} + \frac{1}{M_f^2}\left[-\frac{M_f\lambda(\mathbf{x}_k)}{(1+M_f\lambda(\mathbf{x}_k))} + \ln(1+M_f\lambda(\mathbf{x}_k))\right] \\
&= f(\mathbf{x}_k) + \frac{1}{M_f^2}\left[-\frac{M_f\lambda(\mathbf{x}_k)}{(1+M_f\lambda(\mathbf{x}_k))}(1+M_f\lambda(\mathbf{x}_k)) + \ln(1+M_f\lambda(\mathbf{x}_k))\right] \\
&= f(\mathbf{x}_k) + \frac{1}{M_f^2}\left[-M_f\lambda(\mathbf{x}_k) + \ln(1+M_f\lambda(\mathbf{x}_k))\right] \\
&\leq f(\mathbf{x}_k) - \frac{1}{M_f^2}\left(\frac{M_f^2\lambda(\mathbf{x}_k)^2}{2(1+M_f\lambda(\mathbf{x}_k))}\right) \\
&= f(\mathbf{x}_k) - \frac{\hat{\eta}_k\lambda(\mathbf{x}_k)^2}{2} \\
&\leq f(\mathbf{x}_k) - \alpha\hat{\eta}_k\lambda(\mathbf{x}_k)^2 \\
&= f(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k)
\end{aligned}$$

where we used the fact that

$$-x + \log(1+x) \leq \frac{-x^2}{2(1+x)}$$

Hence, the stepsize is bounded below by

$$\eta_k \geq \frac{\beta}{1+M_f\lambda(\mathbf{x}_k)}$$

Now note that the ouput of our line-search satisfies

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

which using the above bound implies that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k) - \alpha\nabla f(\mathbf{x}_k)^\top(\mathbf{x}_{k+1} - \mathbf{x}_k) = f(\mathbf{x}_k) - \frac{\alpha\beta}{1+M_f\lambda(\mathbf{x}_k)}\lambda(\mathbf{x}_k)^2$$

Hence, the decrease amount in the damped phase is at least

$$\gamma = \frac{\alpha\beta((1-2\alpha)/4M_f)^2}{1+M_f((1-2\alpha)/4M_f)}$$

5 Local analysis

If $\lambda(\mathbf{x}_k)M_f \leq (1-2\alpha)/2$ then for stepsize $\eta_k = 1$ we have

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) - \lambda(\mathbf{x}_k)^2 - \frac{1}{M_f^2}[M_f\lambda(\mathbf{x}_k) - \ln(1 - M_f\lambda(\mathbf{x}_k))] \\
&\leq f(\mathbf{x}_k) - \lambda(\mathbf{x}_k)^2 + \frac{1}{M_f^2}\left[\frac{M_f^2\lambda(\mathbf{x}_k)^2}{2} + M_f^3\lambda(\mathbf{x}_k)^3\right] \\
&= f(\mathbf{x}_k) - \frac{\lambda(\mathbf{x}_k)^2}{2} + M_f\lambda(\mathbf{x}_k)^3 \\
&= f(\mathbf{x}_k) - \lambda(\mathbf{x}_k)^2(\frac{1}{2} - M_f\lambda(\mathbf{x}_k)) \\
&\leq f(\mathbf{x}_k) - \alpha\lambda(\mathbf{x}_k)^2
\end{aligned}$$

Hence, BTLS selects $\eta_k = 1$ for $\lambda(\mathbf{x}_k)M_f < (1-2\alpha)/4$.

It can be further be shown that for $\lambda(\mathbf{x}_k)M_f \leq 1$ we have

$$\lambda(\mathbf{x}_{k+1}) \leq \frac{M_f\lambda(\mathbf{x}_k)^2}{(1 - M_f\lambda(\mathbf{x}_k))^2}$$

Lemma 4. *Consider a self-concordant function f. Show that if $\lambda(\mathbf{x}_k)M_f \leq 1$, then*

$$\lambda(\mathbf{x}_{k+1}) \leq \frac{M_f\lambda(\mathbf{x}_k)^2}{(1 - M_f\lambda(\mathbf{x}_k))^2}$$

Proof. To prove this note that

$$\lambda(\mathbf{x}_{k+1}) = \sqrt{\mathbf{g}_{k+1}^\top \mathbf{H}_{k+1}^{-1}\mathbf{g}_{k+1}} \leq \frac{1}{(1 - M_f r)}\ \sqrt{\mathbf{g}_{k+1}^\top \mathbf{H}_k^{-1}\mathbf{g}_{k+1}} = \frac{1}{(1 - M_f\lambda(\mathbf{x}_k))}\ \sqrt{\mathbf{g}_{k+1}\mathbf{H}_k^{-1}\mathbf{g}_{k+1}}$$

Now we have

$$\begin{aligned}
\mathbf{g}_{k+1}^\top \mathbf{H}_k^{-1}\mathbf{g}_{k+1} &= (\mathbf{g}_k + \mathbf{G}(\mathbf{x}_{k+1} - \mathbf{x}_k))^\top \mathbf{H}_k^{-1}(\mathbf{g}_k + \mathbf{G}(\mathbf{x}_{k+1} - \mathbf{x}_k)) \\
&= (\mathbf{g}_k - \mathbf{G}\mathbf{H}_k^{-1}\mathbf{g}_k)^\top \mathbf{H}_k^{-1}(\mathbf{g}_k - \mathbf{G}\mathbf{H}_k^{-1}\mathbf{g}_k) \\
&= \mathbf{g}_k^\top(\mathbf{I} - \mathbf{G}\mathbf{H}_k^{-1})^\top \mathbf{H}_k^{-1}(\mathbf{I} - \mathbf{G}\mathbf{H}_k^{-1})\mathbf{g}_k \\
&= \mathbf{g}_k^\top(\mathbf{I} - \mathbf{H}_k^{-1}\mathbf{G})\mathbf{H}_k^{-1}(\mathbf{I} - \mathbf{G}\mathbf{H}_k^{-1})\mathbf{g}_k \\
&= \mathbf{g}_k^\top \mathbf{H}_k^{-1/2}(\mathbf{H}_k^{1/2} - \mathbf{H}_k^{-1/2}\mathbf{G})\mathbf{H}_k^{-1}(\mathbf{H}_k^{1/2} - \mathbf{G}\mathbf{H}_k^{-1/2})\mathbf{H}_k^{-1/2}\mathbf{g}_k \\
&= \mathbf{g}_k^\top \mathbf{H}_k^{-1/2}(\mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2})(\mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2})\mathbf{H}_k^{-1/2}\mathbf{g}_k \\
&= \|(\mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2})\mathbf{H}_k^{-1/2}\mathbf{g}_k\|^2 \\
&\leq \|(\mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2})\|^2\|\mathbf{H}_k^{-1/2}\mathbf{g}_k\|^2
\end{aligned}$$

Further, we have

$$1 - M_f r + \frac{1}{3}M_f^2 r^2\mathbf{I} \preceq \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2} \preceq \frac{1}{1 - M_f r}\mathbf{I}$$

Hence,

$$(M_f r - \frac{1}{3}M_f^2 r^2)\mathbf{I} \succeq \mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2} \succeq 1 - \frac{1}{1 - M_f r}\mathbf{I}$$

which implies

$$M_f r\mathbf{I} \succeq \mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2} \succeq -\frac{M_f r}{1 - M_f r}\mathbf{I}$$

Using these bounds we obtain that

$$\|\mathbf{I} - \mathbf{H}_k^{-1/2}\mathbf{G}\mathbf{H}_k^{-1/2}\| \leq \max\{M_f r, \frac{M_f r}{1 - M_f r}\} \leq \frac{M_f r}{1 - M_f r}$$

and the claim follows using the fact that

$$\lambda(\mathbf{x}_{k+1}) \leq \frac{1}{(1 - M_f\lambda(\mathbf{x}_k))} \sqrt{\mathbf{g}_{k+1}\mathbf{H}_k^{-1}\mathbf{g}_{k+1}} \leq \frac{\lambda(\mathbf{x}_k)}{(1 - M_f\lambda(\mathbf{x}_k))} \frac{M_f r}{1 - M_f r} = \frac{M_f\lambda(\mathbf{x}_k)^2}{(1 - M_f\lambda(\mathbf{x}_k))^2}$$

where the last equality follows from the fact that $r = \lambda(\mathbf{x}_k)$. □

Now consider the above result for the case that $\lambda(\mathbf{x}_k)M_f \leq (1-2\alpha)/4$ which implies that $\lambda(\mathbf{x}_k)M_f < 1/4$. In this case, we have

$$\lambda(\mathbf{x}_{k+1}) \leq 2M_f\lambda(\mathbf{x}_k)^2$$

which implies that

$$2M_f\lambda(\mathbf{x}_{k+1}) \leq (2M_f\lambda(\mathbf{x}_k))^2$$

Hence, after ℓ steps in the local phase we have

$$2M_f\lambda(\mathbf{x}_\ell) \leq (2M_f\lambda(\mathbf{x}_{k_0}))^{2^{\ell-k_0}} \leq (0.5)^{2^{\ell-k_0}}$$

6 Overall Cost

The above analysis shows that the overall complexity of Newton's method in this case is

$$\mathcal{O}((1 + M_f^2)(f(\mathbf{x}_0) - f^*) + \log\log(1/\epsilon))$$