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

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

Lecture 17

Caramanis & Mokhtari Monday, March 28, 2022

Goal: In this lecture we talk about Newton's method and its convergence properties.

1 Introduction to Newton's Method

The convergence of gradient descent depends inherently on the condition number; a change of coordinates can improve the condition number, and hence we should do a change of coordinates at each step. There are indeed several motivations for the Newton step. Before we discuss these, we give the basic definition of the Newton step.

Definition 1. *For f having positive definite Hessian $\nabla^2 f(\mathbf{x}) \succ 0$, the Newton updating rule with step size η is defined as,*

$$\mathbf{x}^+ = \mathbf{x} + \eta \Delta \mathbf{x}_{\text{nt}} = \mathbf{x} - \eta \, \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}), \tag{1}$$

where, $\Delta \mathbf{x}_{\text{nt}}$ is called the Newton step.

Note that since $\nabla^2 f(\mathbf{x}) \succ 0$ we have,

$$\nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\text{nt}} = -\nabla f(\mathbf{x})^T \nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x}) < 0$$

always holds for $\nabla f(\mathbf{x}) \neq 0$, so this is a descent direction.
There are various ways to interpret this choice of updating rule.

Minimizer of Quadratic Approximation

Consider a quadratic approximation of f around $\mathbf{x}$,

$$\tilde{f}(\mathbf{x} + \mathbf{v}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{v} + \frac{1}{2} \mathbf{v}^\top \nabla^2 f(\mathbf{x}) \mathbf{v}. \tag{2}$$

This quadratic function is minimized at $\mathbf{v}^\star = -\nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$. Note that if f is quadratic, this approximation is exact and $\mathbf{x} + \mathbf{v}^\star$ is the exact minimizer of f.

Linear Approximation of Gradient around x

Consider a linear approximation of $\nabla f(\mathbf{x} + \mathbf{v})$,

$$\nabla f(\mathbf{x} + \mathbf{v}) \simeq \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x}) \mathbf{v}. \tag{3}$$

The Newton updating rule is obtained by setting the right hand side to 0, which is an approximation to the optimality condition $\nabla f(\mathbf{x}^\star) = 0$.

2 Convergence of Newton's Method

We make two major assumptions in this analysis:

1. f is twice differentiable, μ-strongly convex, and L_1-smooth, such that $\mu I \preceq \nabla^2 f(\mathbf{x}) \preceq L_1 I$.
2. $\nabla^2 f(\mathbf{x})$ is Lipschitz continuous with constant $L_2 > 0$, i.e.

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L_2 \|\mathbf{x} - \mathbf{y}\|_2 \quad \text{for all } \mathbf{x}, \mathbf{y}. \tag{4}$$

The first assumption is the standard assumption we made for gradient descent, namely, that our function f is smooth (in the sense that it has an upper bound on curvature), and it is strongly convex (it has a lower bound on curvature). The assumption also requires that f be twice differentiable, for the definition to make sense.

Beyond the assumptions for gradient descent, we also need some control on the smoothness of the Hessian. This is given in the second assumption. Note that the norm on the left is the spectral norm, defined as the largest singular value (which coincides with the largest eigenvalue, for positive semidefinite matrices). L_2 can be interpreted as a bound on the third derivative of f. The smaller L_2 is, the better f can be approximated by a quadratic function. Since each step of Newton's method minimizes a quadratic approximation of f, the performance of Newton's method works best for functions with small L_2.

Lemma 1. *Consider a twice differentiable function $f : \mathbb{R}^n \to R$. Further assume that $\nabla^2 f(\mathbf{x})$ is Lipschitz continuous with constant $L_2 > 0$. Then, we have*

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2} \nabla f(\mathbf{x})^\top \nabla^2 f(\mathbf{x}) \nabla f(\mathbf{x}) + \frac{L_2}{6} \|\mathbf{y} - \mathbf{x}\|^3$$

Proof. Exercise. □

Lemma 2. *Consider a twice differentiable function $f : \mathbb{R}^n \to R$. Further assume that $\nabla^2 f(\mathbf{x})$ is Lipschitz continuous with constant $L_2 > 0$. Then, we have*

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

Proof. According to the fundamental theorem of calculus we have

$$\begin{aligned}
\nabla f(\mathbf{y}) &= \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f((1-\alpha)\mathbf{x} + \alpha \mathbf{y})(\mathbf{y} - \mathbf{x}) d\alpha \\
&= \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \int_0^1 \big(\nabla^2 f((1-\alpha)\mathbf{x} + \alpha \mathbf{y}) - \nabla^2 f(\mathbf{x})\big)(\mathbf{y} - \mathbf{x}) d\alpha
\end{aligned}$$

and therefore we have

$$\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) = \int_0^1 \big(\nabla^2 f((1-\alpha)\mathbf{x} + \alpha \mathbf{y}) - \nabla^2 f(\mathbf{x})\big)(\mathbf{y} - \mathbf{x}) d\alpha$$

By computing the norm of both sides and using L-smoothness of the Hessians we have

$$\begin{aligned}
\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}) - \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x})\| &= \left\| \int_0^1 \big(\nabla^2 f((1-\alpha)\mathbf{x} + \alpha \mathbf{y}) - \nabla^2 f(\mathbf{x})\big)(\mathbf{y} - \mathbf{x}) d\alpha \right\| \\
&\leq \int_0^1 \left\| \nabla^2 f((1-\alpha)\mathbf{x} + \alpha \mathbf{y}) - \nabla^2 f(\mathbf{x}) \right\| \|\mathbf{y} - \mathbf{x}\| d\alpha \\
&\leq \int_0^1 L_2 \alpha \|\mathbf{y} - \mathbf{x}\|^2 d\alpha = \frac{L_2}{2} \|\mathbf{y} - \mathbf{x}\|^2
\end{aligned}$$

□

For notational convenience we denote $\mathbf{g} = \nabla f(\mathbf{x})$ and $\mathbf{H} = \nabla^2 f(\mathbf{x})$ from this point on.
Our main result of this lecture is Theorem 1. We devote the rest of the lecture trying to understand and prove this theorem. Before stating the theorem, let us first recall Backtracking Line Search (BTLS). With BTLS, first α and β are chosen such that $0 < \alpha < \frac{1}{2}$ and $0 < \beta < 1$, starting with $\eta = 1$, repeat

while true
  if $f(\mathbf{x} + \eta\Delta\mathbf{x}) \leq f(\mathbf{x}) + \alpha\eta\mathbf{g}^\top\Delta\mathbf{x}$
    exit
  else
    $\eta \leftarrow \beta\eta$
  end
end

We are now ready to state and prove the theorem.

Theorem 1. *Consider the update of Newton's method using BTLS with parameters $0 < \alpha < 0.5$ and $0 < \beta < 1$. Further, consider the definitions $\zeta := \min\{1, 3(1-2\alpha)\}\frac{\mu^2}{L_2}$ and $\gamma := \alpha\beta\eta^2\frac{\mu}{L_1^2}$. Then we have*

(a). Damped Newton Phase: If $\|\nabla f(\mathbf{x})\|_2 \geq \zeta$ then $f(\mathbf{x}^+) - f(\mathbf{x}) \leq -\gamma$.

(b). Quadratic Phase: If $\|\nabla f(\mathbf{x})\|_2 < \zeta$ then BTLS $\eta = 1$ and

$$\frac{L_2}{2\mu^2}\|\nabla f(\mathbf{x}^+)\|_2 \leq \left(\frac{L_2}{2\mu^2}\|\nabla f(\mathbf{x})\|_2\right)^2. \tag{5}$$

2.1 Convergence Proof

For readability of the proof, we divide the proof into lemmas to emphasize the flow of the proof and not get lost in the details of the derivation.

Lemma 3. *With the assumptions in part (a), $\eta = \frac{\mu}{L_1}$ satisfies BTLS exit condition, i.e., $f(\mathbf{x} + \eta\Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) + \alpha\eta\mathbf{g}^\top\Delta\mathbf{x}_{\text{nt}}$.*

Proof.

$$f(\mathbf{x}^+) = f(\mathbf{x} - \eta\mathbf{H}^{-1}\mathbf{g}) \tag{6}$$

$$\leq f(\mathbf{x}) - \eta\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{g} + \frac{L_1}{2}\eta^2\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{g} \tag{7}$$

Note that[1],

$$\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{H}^{-1}\mathbf{g} = \mathbf{g}^\top\mathbf{H}^{-1/2}\mathbf{H}^{-1}\mathbf{H}^{-1/2}\mathbf{g} \leq \frac{1}{\mu}\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{g} \tag{8}$$

[1]Recall the definition of square root of a matrix. If A is positive definite then we can write $A = U\Lambda U^T$, where U is unitary and Λ is diagonal, and $A^{1/2} = U\Lambda^{1/2}U^T$.

Thus,

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \eta \mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} + \frac{L_1}{2\mu}\eta^2 \mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} \tag{9}$$

Setting $\eta = \frac{\mu}{L_1}$,

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \frac{\mu}{2L_1}\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} \tag{10}$$

This satisfies the exit condition for $\eta = \frac{\mu}{L_1}$, $f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha \frac{\mu}{L_1}\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g}$ since $\alpha < 1/2$. □

This result shows that if we are in phase (a) then $\eta \geq \beta\mu/L_1$.

Lemma 4. *With the assumptions in part (b), $\eta = 1$ satisfies BTLS exit condition.*

Proof. In this proof we find $\alpha < \frac{1}{2}$ such that $\eta = 1$ satisfies BTLS exit condition. Our goal is to find α such that,

$$f(\mathbf{x} + \Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) + \alpha \mathbf{g}^\top \Delta\mathbf{x}_{\text{nt}}. \tag{11}$$

For notational convenience we denote

$$\lambda(\mathbf{x}) = (\Delta\mathbf{x}_{\text{nt}}^T \mathbf{H} \Delta\mathbf{x}_{\text{nt}})^{1/2} = (\mathbf{g}^\top \mathbf{H}^{-1} g)^{1/2} \tag{12}$$

which is known as the Newton decrement at $\mathbf{x}$. The second equality follows because $\Delta\mathbf{x}_{\text{nt}} = -\mathbf{H}^{-1}\mathbf{g}$. By Lemma **??** we know that if we follow the update of Newton's method with stepsize $\eta = 1$, i.e., $\mathbf{x}^+ = \mathbf{x} - \mathbf{H}^{-1}\mathbf{g}$, we have

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} + \frac{1}{2}\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} + \frac{L_2}{6}\|\mathbf{H}^{-1}\mathbf{g}\|^3$$

which can be upper bounded by

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} + \frac{1}{2}\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g} + \frac{L_2}{6\mu^{3/2}}\|\mathbf{H}^{-1/2}\mathbf{g}\|^3$$

and simplified as

$$\begin{aligned} f(\mathbf{x} + \Delta\mathbf{x}_{\text{nt}}) &\leq f(\mathbf{x}) - \frac{1}{2}\lambda^2(\mathbf{x}) + \frac{L_2}{6\mu^{3/2}}\lambda^3(\mathbf{x}) \\ &= f(\mathbf{x}) - \lambda^2(\mathbf{x})\left(\frac{1}{2} - \frac{L_2\lambda(\mathbf{x})}{6\mu^{3/2}}\right) \\ &= f(\mathbf{x}) + \mathbf{g}^\top \Delta\mathbf{x}_{\text{nt}}\left(\frac{1}{2} - \frac{L_2\lambda(\mathbf{x})}{6\mu^{3/2}}\right) \end{aligned}$$

Note that since $\zeta \leq 3(1-2\alpha)\frac{\mu^2}{L_2}$ we can show that

$$\lambda(\mathbf{x}) = (\mathbf{g}^\top \mathbf{H}^{-1}\mathbf{g})^{1/2} \leq \frac{1}{\mu^{1/2}}\|\mathbf{g}\|_2 < \frac{1}{\mu^{1/2}}\zeta \leq 3(1-2\alpha)\frac{\mu^{3/2}}{L_2}.$$

By regrouping the terms we can show that

$$\alpha < \frac{1}{2} - \frac{L_2\lambda(\mathbf{x})}{6\mu^{3/2}}$$

and hence we have

$$f(\mathbf{x} + \Delta\mathbf{x}_{\text{nt}}) \leq f(\mathbf{x}) + \alpha\mathbf{g}^\top \Delta\mathbf{x}_{\text{nt}}$$

then $\eta = 1$ satisfies BTLS exit condition.

□

Proof of Theorem 1

Theorem 1 part (a). Using Lemma 3, with $\eta \geq \beta\frac{\mu}{L_1}$ and substituting $\Delta\mathbf{x}_{\text{nt}} = -\mathbf{H}^{-1}\mathbf{g}$, we have

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha\beta\frac{\mu}{L_1}\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{g}. \tag{13}$$

By strong convexity $\mathbf{H} \preceq L_1\mathbf{I}$, so $\mathbf{H}^{-1} \succeq \frac{1}{L_1}\mathbf{I}$, and we have

$$\mathbf{g}^\top\mathbf{H}^{-1}\mathbf{g} \geq \frac{1}{L_1}\|\mathbf{g}\|_2^2, \tag{14}$$

therefore,

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) - \alpha\beta\frac{\mu}{L_1}\frac{1}{L_1}\|\mathbf{g}\|_2^2 \tag{15}$$

$$f(\mathbf{x}^+) - f(\mathbf{x}) \leq -\underbrace{\alpha\beta\frac{\mu}{L_1^2}\zeta^2}_{\gamma} \tag{16}$$

where the last inequality follows because $\|\mathbf{g}\|_2 \geq \zeta$. □

Theorem 1 part (b). Using Lemma 4, we can set $\eta = 1$. By using Lemma 2 we can write that

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

which can be simplified as

$$\|\nabla f(\mathbf{x}^+)\| \leq \frac{L_2}{2}\|\mathbf{H}^{-1}\mathbf{g}\|^2$$

Also, $\mathbf{H} \succeq \mu\mathbf{I}$ so $\mathbf{H}^{-1} \preceq \frac{1}{\mu}\mathbf{I}$ and,

$$\|\mathbf{H}^{-1}\mathbf{g}\|_2^2 \leq \frac{1}{\mu^2}\|\mathbf{g}\|_2^2. \tag{17}$$

Substituting this,

$$\|\nabla f(\mathbf{x}^+)\|_2 \leq \frac{L_2}{2\mu^2}\|\mathbf{g}\|_2^2, \tag{18}$$

and multiplying both sides by $\frac{L_2}{2\mu^2}$ we obtain the result stated in the theorem. □

Implication of (a)

In the damped Newton phase, f decreases by at least γ at each iteration, there the total of iterations in this phase cannot exceed,

$$\frac{f(\mathbf{x}_0) - f^\star}{\gamma}$$

since otherwise $f(\mathbf{x})$ would be less than $f^\star$, which contradicts the optimality of $f^\star$. In other words, the quadratic phase starts after at most $\frac{f(\mathbf{x}_0)-f^\star}{\gamma}$ iterations.

Implication of (b)

Let k be the first iteration in which $\|\mathbf{g}\| < \zeta$. And let $\ell \geq 0$ be the number of iterations after k. For simplicity, let us define:

$$a_\ell = \frac{L_2}{2\mu^2}\|\nabla f(\mathbf{x}^{(k+\ell-1)})\|_2 \tag{19}$$

First, let us establish a bound on a_1. In the quadratic phase, since $\|\nabla f(\mathbf{x}_k)\|_2 < \zeta$ and $\zeta < \frac{\mu^2}{L_2}$ by assumption, we have

$$\frac{L_2}{2\mu^2}\|\nabla f(\mathbf{x}_k)\|_2 < \frac{L_2}{2\mu^2}\eta < \frac{1}{2}.$$

Thus, $a_1 < \frac{1}{2}$. Now from (b) of the theorem, we also have that $a_{\ell+1} \leq a_\ell^2$. Therefore, we have the following sequence:

$$a_\ell \leq (a_{\ell-1})^2 \leq (a_{\ell-2})^{2^2} \leq (a_{\ell-3})^{2^3} \leq \cdots \leq (a_1)^{2^{\ell-1}} \implies a_\ell \leq (a_1)^{2^{\ell-1}}$$

$$\implies a_\ell \leq \left(\frac{1}{2}\right)^{2^{\ell-1}}$$

$$\implies \frac{L}{2\mu^2}\|\nabla f(\mathbf{x}_\ell)\|_2 \leq \left(\frac{1}{2}\right)^{2^{\ell-1}}$$

$$\implies \|\nabla f(\mathbf{x}_\ell)\|_2 \leq \frac{2\mu^2}{L}\left(\frac{1}{2}\right)^{2^{\ell-1}}$$

For strongly convex functions, we also have

$$f(\mathbf{x}_\ell) - f^\star \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x}_\ell)\|_2^2 \tag{20}$$

$$\leq \frac{1}{2\mu}\left(\frac{2\mu^2}{L_2}\left(\frac{1}{2}\right)^{2^{\ell-1}}\right)^2 \tag{21}$$

$$\leq \frac{2\mu^3}{L_2^2}\left(\frac{1}{2}\right)^{2^{\ell-1}} \tag{22}$$

thus, $f(\mathbf{x}) \to f^*$ quadratically.

Therefore, if we want $a_\ell \leq \epsilon$, we only need the following number of iterations:

$$\begin{aligned}
(a_1)^{2^{\ell-1}} &\leq \epsilon \\
2^{\ell-1} \log a_1 &\leq \log \epsilon \\
2^{\ell-1} &\geq \text{constant} \times \log \epsilon \\
\ell - 1 &\geq \log \log \epsilon + \text{constant}.
\end{aligned}$$

Hence, the number of iterations k required for Newton'e method to find a solution satisfying the condition $f(\mathbf{x}_k) - f^* \leq \epsilon$ is

$$k = \frac{L_1^2 L_2^2 (f(\mathbf{x}_0) - f^*)}{\mu^5 \alpha \beta \min\{1, 9(1-2\alpha)^2\}} + \log \log \left(\frac{L_2^2}{2\mu^3 \epsilon} \right)$$