MIT 6.7220/15.084 — Nonlinear Optimization Tue, Apr 9th 2024

Lecture 12
Hessians, preconditioning, and Newton's method

Instructor: Prof. Gabriele Farina (✉ gfarina@mit.edu)*

So far, we have been concerned with *first-order* methods, that is, those optimization methods that use gradient information. With today's lecture, we will start discussing *second-order* methods, which use not only the gradient but also the *Hessian* (second-order derivative) of the function to be optimized. For that, we will now restrict our attention to optimization problems of the form

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{s.t.} \quad & x \in \mathbb{R}^n \end{aligned}$$

where $f(x)$ is a *twice-differentiable* function.

1 From first-order to second-order Taylor approximations

As we mentioned in Lecture 7 when introducing the gradient descent algorithm, a fundamental idea for constructing optimization algorithms is to approximate the function to be optimized by a simpler function that is easier to minimize. In the case of gradient descent, we picked a direction of movement based on the minimum of the *first-order* Taylor expansion of the objective function around the current point.

In the case of twice-differentiable functions, it seems natural to wonder what happens if we instead were to pick the direction of movement by looking instead at the *second-order* Taylor expansion of the objective, which is a more faithful approximation of the function.

The second-order Taylor expansion of a function $f(x)$ around a point x_t is given by

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2} \langle x - x_t, \nabla^2 f(x_t)(x - x_t) \rangle$$

where $\nabla^2 f(x_t)$ is the Hessian matrix of f at x_t. The minimum of $f(x)$ can be found in closed form by setting the gradient (with respect to x) of the above expression to zero, which gives

$$\nabla f(x_t) + \nabla^2 f(x_t)(x - x_t) = 0 \qquad \Longrightarrow \qquad x = x_t - [\nabla^2 f(x_t)]^{-1} \nabla f(x_t).$$

So, we find that by moving from first-order to second-order Taylor approximation, and assuming that $\nabla^2 f(x_t)$ is invertible, the natural direction of descent changes from

$$\underbrace{d = -\nabla f(x_t)}_{\substack{\text{using first-order} \\ \text{Taylor approximation}}} \qquad \text{to} \qquad \underbrace{d = -[\nabla^2 f(x_t)]^{-1} \nabla f(x_t)}_{\text{using second-order Taylor approximation}} .$$

*These notes are class material that has not undergone formal peer review. The TAs and I are grateful for any reports of typos.

1.1 Second-order direction as an example of preconditioning

The second-order direction of descent is obtained by multiplying the negative gradient (that is, the first-order direction of descent) by the inverse of the Hessian matrix. This operation is known as *preconditioning* the gradient by the Hessian.

The Hessian preconditioning of the gradient direction has the effect of making the optimization problem affinely invariant. In other words, if we apply an affine transformation to the objective function and the initial point, the second-order direction of descent will also be automatically transformed.

To demonstrate this property, consider the optimization problem

$$\begin{aligned} &\min_x \ f(x) \\ &\text{s.t.} \ \ x \in \mathbb{R}^n \end{aligned}$$

and suppose that the coordinates of x have been reparametrized via a new coordinate system y where the correspondence to x is given by $x \leftrightarrow Ay + b$.

In the coordinate system of y, the function being minimized is $g(y) \coloneqq f(Ay + b)$; so, the gradient and Hessian of g can be computed as:

$$\nabla_y g(y) = A^\top \nabla_x f(x), \qquad \nabla_y^2 g(y) = A^\top \nabla_x^2 f(x) A$$

So,

$$\begin{aligned} \underbrace{-\left[\nabla_y^2 g(y)\right]^{-1} \nabla_y g(y)}_{d_y} &= -\left[A^\top \nabla_x^2 f(x) A\right]^{-1} \left(A^\top \nabla_x f(x)\right) \\ &= -A^{-1} \left[\nabla_x^2 f(x)\right]^{-1} A^{-\top} A^\top \nabla_x f(x) \\ &= A^{-1} \cdot \underbrace{\left(-\left[\nabla_x^2 f(x)\right]^{-1} \nabla_x f(x)\right)}_{d_x}, \end{aligned}$$

This implies that the second-order directions of descent measured with respect to y and x satisfy

$$d_x \leftrightarrow A d_y,$$

which mimics the correspondence $x \leftrightarrow Ay + b$. So, for example, the update $x' = x - \eta d_x$ in the x coordinate system corresponds to the update

$$x' = (Ay + b) - \eta A d_y = A\big(y - \eta d_y\big) + b = Ay' + b$$

in the y coordinate system, preserving the correspondence at all steps. The same property does not hold for gradient descent. We will have a more in-depth look into preconditioning in the next Lecture.

2 From gradient descent to Newton's method

Having established the second-order direction of descent, we can define the natural generalization of the gradient descent algorithm to the second-order setting. This algorithm, which takes the name *damped Newton's method*, is given by the update rule

$$\boxed{x_{t+1} = x_t - \eta \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t)} \,. \tag{1}$$

For the choice $\eta = 1$, which corresponds to setting x_{t+1} to be the minimum of the second-order approximation centered at x_t at each iteration, this algorithm is known simply as *Newton's method*.

2.1 Failure mode: lack of curvature

Given that the update rule of Newton's method involves the inverse of the Hessian, a natural concern is what happens when the Hessian is singular or near-singular (for example, when one eigenvalue is extremely close to zero). Such a situation corresponds to the case where the function has very little curvature, and the second-order approximation is not a good approximation of the function. We now show that these concerns are well-founded by considering a function whose curvature away from the minimum decays extremely fast.

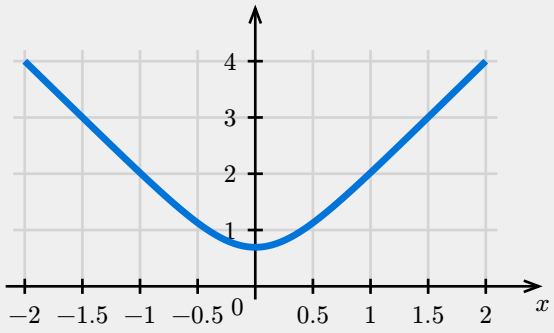
Example 2.1. Consider the function

$$f(x) = \log(e^{2x} + e^{-2x}),$$

plotted on the right, whose gradient and Hessian are respectively computed as

$$\nabla f(x) = 2 \cdot \frac{e^{4x} - 1}{e^{4x} + 1},$$

$$\nabla^2 f(x) = 16 \cdot \frac{e^{4x}}{(e^{4x} + 1)^2}.$$



The two tables below show the first 10 iterates of Newton's method and gradient descent when applied to $f(x)$ starting at two close initial points: $x_0 = 0.5$ on the left, and $x_0 = 0.7$ on the right. As you can see, the behavior of Newton's method is very different: while it converges extremely quickly to the minimum when starting at $x_0 = 0.5$, it diverges when starting at $x_0 = 0.7$.

	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.5000**	**0.5000**
$t = 1$	−0.4067	0.3477
$t = 2$	0.2047	0.2274
$t = 3$	−0.0237	0.1422
$t = 4$	3.53×10^{-5}	0.0868
$t = 5$	-1.17×10^{-13}	0.0524
$t = 6$	-1.14×10^{-17}	0.0315
$t = 7$	0.0000	0.0189
$t = 8$	0.0000	0.0114
$t = 9$	0.0000	0.0068

	Newton's method	GD ($\eta = 0.1$)
$t = 0$	**0.7000**	**0.7000**
$t = 1$	−1.3480	0.5229
$t = 2$	26.1045	0.3669
$t = 3$	-2.79×10^{44}	0.2418
$t = 4$	diverged	0.1520
$t = 5$	diverged	0.0930
$t = 6$	diverged	0.0562
$t = 7$	diverged	0.0338
$t = 8$	diverged	0.0203
$t = 9$	diverged	0.0122

In the next section, we will analyze the convergence properties of Newton's method formally.

3 Analysis of Newton's method

Example 2.1 shows that Newton's method breaks down in the absence of sufficient curvature. However, it also showed that the method can be extremely efficient when the starting point is close to the minimum and the minimum has enough curvature. In this section, we formalize these positive observations quantitatively.

3.1 A key lemma

The analysis of Newton's method and damped Newton's method hinges on the following lemma, which relates the geometric shrinking in the distance to optimality of the iterates to the spectral norm of a specific matrix that depends on the Hessian of the function.

Theorem 3.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with invertible Hessian, and let $x_\star$ be a local minimum of f. The distance to optimality of the iterates x_t generated by the damped Newton's method with stepsize $\eta > 0$ satisfy

$$x_{t+1} - x_\star = (I - \eta H_t)(x_t - x_\star),$$

where

$$H_t \coloneqq \left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star))\, \mathrm{d}\lambda.$$

Proof. The fundamental idea of the proof is to write the second-order direction as a function of the Hessian matrices on the segment connecting $x_\star$ to x_t.

In particular, we have

$$\begin{aligned}
\left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t) &= \left[\nabla^2 f(x_t)\right]^{-1} (\nabla f(x_t) - \nabla f(x_\star)) && \text{(since } \nabla f(x_\star) = 0\text{)} \\
&= \left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star))(x_t - x_\star)\, \mathrm{d}\lambda \\
&= \left(\left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x_t - x_\star))\, \mathrm{d}\lambda \right) (x_t - x_\star) \\
&= H_t (x_t - x_\star).
\end{aligned}$$

Hence, substituting this expression in the update rule of damped Newton's method, we find

$$x_{t+1} - x_\star = (x_t - x_\star) - \eta \left[\nabla^2 f(x_t)\right]^{-1} \nabla f(x_t) = (I - \eta H_t)(x_t - x_\star),$$

as we wanted to show. □

The previous lemma shows that as long as the spectral norm of $I - \eta H_t$ is less than 1 (that is, the maximum absolute value of any eigenvalue of $I - \eta H_t$ is less than 1), the distance to optimality of the iterates x_t generated by Newton's method will decay exponentially fast. We will leverage this result in the next two subsection to give local and global convergence guarantees under different hypotheses.

3.2 First corollary: Local convergence guarantees

As a first corollary of the previous lemma, we can derive a convergence guarantee for Newton's method when starting from a point that is "close enough" to a minimum with sufficient curvature. This type of guarantee is known as a *local convergence guarantee*, as it only applies to points that are within a certain distance from the solution. An empirical illustration of this guarantee is shown in Example 2.1 when starting from the initial point $x_0 = 0.5$ (left table of the example).

In particular, let $x_\star$ be a local minimum of f with strong curvature, that is, a point such that

$$\nabla f(x_\star) = 0, \qquad \text{and} \qquad \nabla^2 f(x_\star) \succcurlyeq \mu I \tag{2}$$

for some $\mu > 0$. Furthermore, assume that f is *smooth*, in the sense that its Hessian is M-Lipschitz continuous, that is

$$\left\|\nabla^2 f(x) - \nabla^2 f(y)\right\|_s \leq M \cdot \|x - y\|_2. \tag{3}$$

Our analysis will rest on the following high-level idea.

Since the Hessian at $x_\star$ is $\succcurlyeq \mu I$ and the Hessian is M-Lipschitz continuous—and therefore cannot change too fast—then we can determine a neighborhood of points "sufficiently close" to $x_\star$ with strong curvature. This will allow us to upper bound the spectral norm of $I - H_t$ and invoke the general result of Theorem 3.1.

We make the above idea formal in the following.

Theorem 3.2. Under the assumptions (2) and (3) above, the spectral norm of the matrix $I - H_t$ induced at time t by the iterate x_t produced by Newton's method satisfies

$$\|I - H_t\|_s \leq \frac{M}{\mu}\|x_t - x_\star\|_2$$

whenever

$$\|x_t - x_\star\|_2 \leq \frac{\mu}{2M}.$$

Proof. The key technique is to leverage the Lipschitz continuity of the Hessian to bound the spectral norm of $I - H_t$. To do that, we can make a difference of Hessians appear in the expression of $I - H_t$ by rewriting the identity matrix as follows:

$$\begin{aligned} I - H_t &= \left(\left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \nabla^2 f(x_t)\, \mathrm{d}\lambda \right) - H_t \\ &= \left[\nabla^2 f(x_t)\right]^{-1} \int_0^1 \Big(\nabla^2 f(x_t) - \nabla^2 f(x_\star + \lambda(x_t - x_\star)) \Big)\, \mathrm{d}\lambda. \end{aligned}$$

So, taking spectral norms on both sides, we find

$$\begin{aligned} \|I - H_t\|_s &\leq \left\| \left[\nabla^2 f(x_t)\right]^{-1} \right\|_s \cdot \int_0^1 \left\| \nabla^2 f(x_t) - \nabla^2 f(x_\star + \lambda(x_t - x_\star)) \right\|_s \mathrm{d}\lambda \\ &\leq \left\| \left[\nabla^2 f(x_t)\right]^{-1} \right\|_s \cdot \int_0^1 M(1-\lambda)\|x_t - x_\star\|_2\, \mathrm{d}\lambda \\ &= \frac{M}{2} \left\| \left[\nabla^2 f(x_t)\right]^{-1} \right\|_s \cdot \|x_t - x_\star\|_2. \end{aligned} \tag{4}$$

To complete the proof, we only need to bound the spectral norm of $\left[\nabla^2 f(x_t)\right]^{-1}$. As mentioned above, we will do so by using the fact that $\nabla^2 f(x_\star) \succcurlyeq \mu I$ and the Hessian changes slowly. In particular, we have

$$\left\|\nabla^2 f(x_t) - \nabla^2 f(x_\star)\right\|_s \leq M\|x_t - x_\star\|_2,$$

which implies that

$$-M\|x_t - x_\star\|_2 I \preccurlyeq \nabla^2 f(x_t) - \nabla^2 f(x_\star) \preccurlyeq M\|x_t - x_\star\|_2 I.$$

In particular,

$$\nabla^2 f(x_t) \succcurlyeq \nabla^2 f(x_\star) - M\|x_t - x_\star\|_2 I \succcurlyeq \big(\mu - M\|x_t - x_\star\|_2\big) I \succcurlyeq \frac{\mu}{2} I,$$

where the last inequality used the hypothesis that $\|x_t - x_\star\|_2 \le \frac{\mu}{2M}$. Hence, $\left\|[\nabla^2 f(x_t)]^{-1}\right\|_s \le \frac{2}{\mu}$, and plugging this bound into (4) yields the statement. □

In particular, since $\|I - H_t\|_s \le 1/2$, then Theorem 3.1 guarantees that the distance to optimality decreases exponentially. In fact, since Theorem 3.2 bounds the spectral norm of $I - H_t$ as a linear function of $\|x_t - x_\star\|_2$, the decrease is doubly exponential. We have just shown the following.

Theorem 3.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with M-Lipschitz continuous Hessian, and let $x_\star$ be a local minimum of f with strong curvature, that is, a point such that

$$\nabla f(x_\star) = 0, \qquad \text{and} \qquad \nabla^2 f(x_\star) \succcurlyeq \mu I$$

for some $\mu > 0$. Then, as long as we start Newton's method from a point x_0 with distance

$$\|x_0 - x_\star\|_2 \le \frac{\mu}{2M}$$

from the local minimum, the distance to optimality of the iterates x_t generated by Newton's method decays as

$$\frac{\|x_{t+1} - x_\star\|_2}{\mu/M} \le \left(\frac{\|x_t - x_\star\|_2}{\mu/M}\right)^2.$$

3.3 Second corollary: Global convergence for functions with strong curvature

In general, obtaining *global* convergence guarantees for Newton's method is a much harder task than obtaining local convergence guarantees. However, we can still obtain global convergence guarantees for Newton's method when the function is both μ-strongly convex and L-smooth, that is,

$$\mu I \preccurlyeq \nabla^2 f(x) \preccurlyeq L I \qquad \forall x \in \mathbb{R}^n.$$

Under these assumptions,

$$\frac{\mu}{L} I \preccurlyeq [\nabla^2 f(x)]^{-1} \int_0^1 \nabla^2 f(x_\star + \lambda(x - x_\star))\, \mathrm{d}\lambda \preccurlyeq \frac{L}{\mu} I$$

so that

$$\left(1 - \eta\frac{L}{\mu}\right) I \preccurlyeq I - \eta H_t \preccurlyeq \left(1 - \eta\frac{\mu}{L}\right) I.$$

Hence, if $\eta \le \frac{\mu}{L}$, we have

$$0 \preccurlyeq I - \eta H_t \preccurlyeq \left(1 - \eta\frac{\mu}{L}\right) I \qquad \Longrightarrow \qquad \|I - \eta H_t\|_s \le 1 - \eta\frac{\mu}{L}$$

and invoking Theorem 3.1 we obtain the following corollary.

Corollary 3.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, μ-strongly convex and L-smooth. Then, the distance to optimality[1] of the iterates x_t generated by damped Newton's method with stepsize $\eta \leq \frac{\mu}{L}$ decays exponentially fast at the rate

$$\|x_{t+1} - x_\star\|_2 \leq \left(1 - \eta\frac{\mu}{L}\right)\|x_t - x_\star\|_2.$$

Remark 3.1. The previous result is made somewhat less appealing by the fact that we already know that gradient descent can reach a similar convergence rate for the same class of smooth and strongly convex function, without even needing to invert the Hessian (see Lecture 7, section on convergence rate for smooth PŁ functions).

4 Further readings

Further material on (damped) Newton's method can be found in Chapter 1.2.4 of Nesterov, Y. [Nes18]'s book. The use of preconditioning in optimization is discussed in Chapter 1.3 of the same book, under the name "variable metric method".

■ **Appendix: Cubic-regularized Newton method.** As we have seen above, when the Hessian matrix is ill conditioned, Newton's method might take huge steps and diverge. For that reason, we were able to show global convergence only in the limited setting of strong curvature (albeit at an extremely fast rate). We will see in a couple of lectures another extremely important function class for which Newton's method can be shown to converge globally: self-concordant functions.

Beyond these isolated classes of benign functions, one might wonder if second-order method can be made more robust to ill-conditioned Hessian matrices. To resolve this question for the positive, Nesterov, Y., & Polyak, B. T. [NP06] proposed a modification of Newton's method that combines the idea of proximal step to control the step size with the second-order information of the Hessian matrix. In particular, in our pursuit of a good second-order method we would like to be able to trade off the following two objectives:

- moving as much as possible in the direction of minimizing the second-order approximation

$$f(x) \approx f(x_t) + \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2}\langle x - x_t, \nabla^2 f(x_t)(x - x_t)\rangle;$$

- staying in a sufficiently small neighborhood of x_t, where the approximation is accurate. The update rule $x_{t+1} = x_t - \eta[\nabla^2 f(x_t)]^{-1}\nabla f(x_t)$ seen in (1) does not guarantee this, as it might take large steps when the Hessian is ill conditioned.

As we saw already in Lecture 9, one way to capture this tension quantitatively is via the notion of *proximal* step. For first-order methods, when using Euclidean distances to measure the step size, this resulted in the choice of next iterate given by

$$x_{t+1} := \arg\min_x \eta\langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2}\|x - x_t\|_2^2,$$

which we saw was exactly equivalent to the gradient descent update rule $x_{t+1} = x_t - \eta\nabla f(x_t)$. For second-order methods, a natural generalization of the approach results in the update

[1] By strong convexity, the function must attain one minimum, which is also unique.

$$\boxed{x_{t+1} \in \arg\min_x \eta\left(\langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2}\langle x - x_t, \nabla^2 f(x_t)(x - x_t)\rangle\right) + \frac{1}{6}\|x - x_t\|_2^3}. \tag{5}$$

The previous update rule is precisely the one proposed by Nesterov, Y., & Polyak, B. T. [NP06]. We will call the resulting method *cubic-regularized Newton's method*. As shown by Nesterov, Y., & Polyak, B. T. [NP06], this method guarantees global convergence for a much broader class of functions than the ones we have seen so far, while still enjoying quadratic convergence rates once the iterates are close enough to the solution. In addition to the original paper by Nesterov, Y., & Polyak, B. T. [NP06], you can find a treatment of cubic-regularized Newton's method in Chapter 4.1 of Nesterov, Y. [Nes18]'s book.

Bibliography

[Nes18] Y. Nesterov, *Lectures on Convex Optimization*. Springer International Publishing, 2018. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-91578-4

[NP06] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, Aug. 2006, doi: 10.1007/s10107-006-0706-8.

5 Appendix: Spectral norms

The spectral norm of a matrix A measures the maximum increase in Euclidean norm that A can produce, that is,

$$\|A\|_s := \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}.$$

(This type of norm, defined as the maximum effect that a function can have on the norm of an input, is known as an *operator norm*.) In other words, if $\|A\|_s \leq k$, then we know that

$$\|Ax\|_2 \leq k\|x\|_2$$

for all vectors x.

The spectral norm of a matrix is closely related to its eigenvalues. In particular, we have the following characterization.

Theorem 5.1. For a symmetric matrix A, the spectral norm is equal to the maximum absolute value of any eigenvalue of A.

Proof. The theorem is straightforward when A is a diagonal matrix, where the maximum is attained by any vector supported only on the coordinate with the maximum absolute eigenvalue.

To show the result in general, we can reduce to the diagonal case by considering the eigendecomposition

$$A = Q^\top \Lambda Q,$$

where Q is an orthogonal matrix (that is, $\|Qv\|_2 = \|Q^\top v\|_2 = \|v\|_2$ for all v) and Λ is a diagonal matrix with the eigenvalues of A on the diagonal (this decomposition exists from the spectral theorem for symmetric real matrices). Then,

$$\max_{x\neq 0}\frac{\|Ax\|_2}{\|x\|_2}=\max_{x\neq 0}\frac{\left\|Q^\top\Lambda Qx\right\|_2}{\|x\|_2}=\max_{x\neq 0}\frac{\|\Lambda Qx\|_2}{\|Qx\|_2},$$

where the last equality used the orthogonality of Q twice (once in the numerator to remove on $Q^\top$, and once in the denominator to add a Q). Since Q is invertible, we can therefore operate a change of variable and write

$$\max_{x\neq 0}\frac{\left\|Q^\top\Lambda Qx\right\|_2}{\|Qx\|_2}=\max_{y\neq 0}\frac{\|\Lambda y\|_2}{\|y\|_2}.$$

The result then follows from the diagonal case. □

An immediate consequence of the previous result is the following:

Corollary 5.1. Let $A\in\mathbb{R}^{n\times n}$ be symmetric. Then, $\|A\|_s\leq k$ if and only if

$$-kI\preccurlyeq A\preccurlyeq kI.$$

Proof. The condition $-kI\preccurlyeq A\preccurlyeq kI$ is equivalent to asking that every eigenvalue of A be in the range $[-k,k]$. The result then follows from the previous theorem. □

Finally, we show that the spectral norm of a product of matrices is submultiplicative. This follows pretty much directly from the definition of the spectral norm.

Theorem 5.2. For any matrices $A,B\in\mathbb{R}^{n\times n}$, we have $\|AB\|_s\leq\|A\|_s\|B\|_s$.

Proof. Let $x\neq 0$ be a vector. Then,

$$\|ABx\|_2\leq\|A\|_s\|Bx\|_2\leq\|A\|_s\|B\|_s\|x\|_2.$$

Dividing both sides by $\|x\|_2$ and taking the maximum over all $x\neq 0$ yields the result. □