# Chapter 6

# Newton's Method

## Contents

## 6.1 1-dimensional case

The Newton method (or Newton-Raphson method, invented by Sir Isaac Newton and formalized by Joseph Raphson) is an iterative method for finding a zero of a differentiable univariate function $f : \mathbb{R} \to \mathbb{R}$. Starting from some number $x_0$, it computes

$$x_{t+1} := x_t - \frac{f(x_t)}{f'(x_t)}, \quad t \geq 0. \tag{6.1}$$

Figure 6.1 shows what happens. $x_{t+1}$ is the point where the tangent line to the graph of $f$ at $(x_t, f(x_t))$ intersects the $x$-axis. In formulas, $x_{t+1}$ is the solution of the linear equation

$$f(x_t) + f'(x_t)(x - x_t) = 0,$$

and this yields the update formula (6.1).

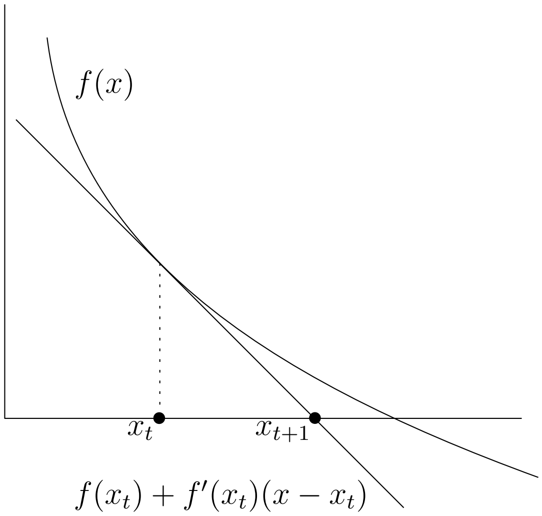

Figure 6.1: One step of Newton's method

The Newton step (6.1) obviously fails if $f'(x_t) = 0$ and may get out of control if $|f'(x_t)|$ is very small. Any theoretical analysis will have to make suitable assumptions to avoid this. But before going into this, we look at Newton's method in a benign case.

Let $f(x) = x^2 - R$, where $R \in \mathbb{R}_+$. $f$ has two zeros, $\sqrt{R}$ and $-\sqrt{R}$. Starting for example at $x_0 = R$, we hope to converge to $\sqrt{R}$ quickly. In this case, (6.1) becomes

$$x_{t+1} = x_t - \frac{x_t^2 - R}{2x_t} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right). \tag{6.2}$$

This is in fact the *Babylonian method* to compute square roots, and here we see that it is just a special case of Newton's method.

Can we prove that we indeed quickly converge to $\sqrt{R}$? What we immediately see from (6.2) is that all iterates will be positive and hence

$$x_{t+1} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) \geq \frac{x_t}{2}.$$

So we cannot be too fast. In order to even get $x_t < 2\sqrt{R}$, we need at least $T \geq \log(R)/2$ steps. It turns out that the Babylonian method starts taking off only when $x_t - \sqrt{R} < 1/2$, say (Exercise 26 asks you to prove that it takes $\mathcal{O}(\log R)$ steps to get there).

To watch takeoff, let us now suppose that $x_0 - \sqrt{R} < 1/2$, so we are starting close to $\sqrt{R}$ already. We rewrite (6.2) as

$$x_{t+1} - \sqrt{R} = \frac{x_t}{2} + \frac{R}{2x_t} - \sqrt{R} = \frac{1}{2x_t}\left(x_t - \sqrt{R}\right)^2. \tag{6.3}$$

Assuming for now that $R \geq 1/4$, all iterates have value at least $\sqrt{R} \geq 1/2$, hence we get

$$x_{t+1} - \sqrt{R} \leq \left(x_t - \sqrt{R}\right)^2.$$

This means that the error goes to 0 *quadratically*, and

$$x_T - \sqrt{R} \leq \left(x_0 - \sqrt{R}\right)^{2^T} < \left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0. \tag{6.4}$$

What does this tell us? In order to get $x_T - \sqrt{R} < \varepsilon$, we only need $T = \log\log(\frac{1}{\varepsilon})$ steps! Hence, it takes a while to get to roughly $\sqrt{R}$, but from there, we achieve high accuracy very fast.

Let us do a concrete example (with IEEE 754 double arithmetic). If $R = 1000$, we need 7 steps to get $x_7 - \sqrt{1000} < 1/2$, and then just 3 more steps to get $x_{10}$ equal to $\sqrt{1000}$ up to the machine precision ($53$ binary digits). In this last phase, we essentially double the number of correct digits in each iteration!

## 6.2 Newton's method for optimization

Suppose we want to find a global minimum $x^\star$ of a differentiable convex function $f : \mathbb{R} \to \mathbb{R}$ (assuming that a global minimum exists). Lemmata 1.12 and Lemma 1.13 guarantee that we can equivalently search for a zero of the derivative $f'$. To do this, we can apply Newton's method if $f$ is *twice* differentiable; the update step then becomes

$$x_{t+1} := x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - f''(x_t)^{-1} f'(x_t), \quad t \geq 0. \tag{6.5}$$

There is no reason to restrict to $d = 1$. Here is Newton's method for minimizing a convex function $f : \mathbb{R}^d \to \mathbb{R}$. We choose $\mathbf{x}_0$ arbitrarily and then iterate:

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t), \quad t \geq 0. \tag{6.6}$$

The update vector $\nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ is the result of a matrix-vector multiplication: we invert the Hessian at $\mathbf{x}_t$ and multiply the result with the gradient at $\mathbf{x}_t$. As before, this fails if the Hessian is not invertible, and may get out of control if the Hessian has small norm.

We have introduced iteration (6.6) simply as a (more or less natural) generalization of (6.5), but there's more to it. If we consider (6.6) as a special case of a general update scheme

$$\mathbf{x}_{t+1} = \mathbf{x}_t - H(\mathbf{x}_t) \nabla f(\mathbf{x}_t),$$

where $H(\mathbf{x}) \in \mathbb{R}^{d \times d}$ is some matrix, then we see that also gradient descent (2.1) is of this form, with $H(\mathbf{x}_t) = \gamma I$. Hence, Newton's method can also be thought of as "adaptive gradient descent" where the adaptation is w.r.t. the local geometry of the function at $\mathbf{x}_t$. Indeed, as we show next, this allows Newton's method to converge on *all* nondegenerate quadratic functions in one step, while gradient descent only does so with the right stepsize on "beautiful" quadratic functions whose sublevel sets are Euclidean balls (Exercise 18).

**Lemma 6.1.** *A* nondegenerate *quadratic function is a function of the form*

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top M \mathbf{x} - \mathbf{q}^\top \mathbf{x} + c,$$

*where $M \in \mathbb{R}^{d \times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d, c \in R$. Let $\mathbf{x}^\star = M^{-1}\mathbf{q}$ be the unique solution of $\nabla f(\mathbf{x}) = \mathbf{0}$ (the unique local minimum if $f$ is convex). With any starting point $\mathbf{x}_0 \in \mathbb{R}^d$, Newton's method (6.6) yields $\mathbf{x}_1 = \mathbf{x}^\star$.*

*Proof.* We have $\nabla f(\mathbf{x}) = M\mathbf{x} - \mathbf{q}$ (this implies $\mathbf{x}^\star = M^{-1}\mathbf{q}$) and $\nabla^2 f(\mathbf{x}) = M$. Hence,

$$\mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1}\nabla f(\mathbf{x}_0) = \mathbf{x}_0 - M^{-1}(M\mathbf{x}_0 - \mathbf{q}) = M^{-1}\mathbf{q} = \mathbf{x}^\star.$$

□

In particular, Newton's method can solve an invertible system $M\mathbf{x} = \mathbf{q}$ of linear equations in one step. But no miracle is happening here, as this step involves the inversion of the matrix $\nabla^2 f(\mathbf{x}_0) = M$.

More generally, the behavior of Newton's method is affine invariant. By this, we mean that it is invariant under any invertible affine transformation, as follows:

**Lemma 6.2** (Exercise 27)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be twice differentiable, $A \in \mathbb{R}^{d\times d}$ an invertible matrix, $\mathbf{b} \in \mathbb{R}^d$. Let $g : \mathbb{R}^d \to \mathbb{R}$ be the (bijective) affine function $g(\mathbf{y}) = A\mathbf{y} + \mathbf{b}, \mathbf{y} \in \mathbb{R}^d$. Finally, let $N_h : \mathbb{R}^d \to \mathbb{R}^d$ denote the Newton step for function $h$, i.e.*

$$N_h(\mathbf{x}) := \mathbf{x} - \nabla^2 h(\mathbf{x})^{-1}\nabla h(\mathbf{x}),$$

*whenever this is defined. Then we have $N_{f\circ g} = g^{-1} \circ N_f \circ g$.*

This says that in order to perform a Newton step for $f \circ g$ on $\mathbf{y}_t$, we can transform $\mathbf{y}_t$ to $\mathbf{x}_t = g(\mathbf{y}_t)$, perform the Newton step for $f$ on $\mathbf{x}$ and transform the result $\mathbf{x}_{t+1}$ back to $\mathbf{y}_{t+1} = g^{-1}(\mathbf{x}_{t+1})$. Another way of saying this is that the following diagram commutes:

$$\begin{tikzcd}
\mathbf{x}_t \arrow[r, "N_f"'] & \mathbf{x}_{t+1} \arrow[d, "g^{-1}"'] \\
\mathbf{y}_t \arrow[u, "g"'] \arrow[r, "N_{f\circ g}"] & \mathbf{y}_{t+1}
\end{tikzcd}$$

Hence, while gradient descent suffers if the coordinates are at very different scales, Newton's method doesn't.

We conclude the general exposition with another interpretation of Newton's method: each step minimizes the local second-order Taylor approximation.

**Lemma 6.3** (Exercise 30)**.** *Let $f$ be convex and twice differentiable at $\mathbf{x}_t \in$ $\mathbf{dom}(f)$, with $\nabla^2 f(\mathbf{x}_t) \succ 0$ being invertible. The vector $\mathbf{x}_{t+1}$ resulting from the Netwon step (6.6) satisfies*

$$\mathbf{x}_{t+1} = \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}-\mathbf{x}_t) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x}-\mathbf{x}_t).$$

## 6.3 Once you're close, you're there...

We will prove a result about Newton's method that may seem rather weak: under suitable conditions, and starting close to the global minimum, we will reach distance at most $\varepsilon$ to the minimum within $\log\log(1/\varepsilon)$ steps. The weak part here is of course not the number of steps $\log\log(1/\varepsilon)$—this is much faster than anything we have seen so far—but the assumption that we are starting close to the minimum already. Under such an assumption, we say that we have a *local convergence* result.

*Global convergence* results that hold for every starting point are unknown for Newton's method as in (6.6). There are some variants of the method for which such results can be proved, most notably the cubic regularization variant of Nesterov and Polyak [NP06]. Weak global convergence results can be obtained by adding a step size to (6.6) and always making only steps that decrease the function value (which may not happen under the full Newton step).

An alternative is to use gradient descent to get us sufficiently close to the global minimum, and then switch to Newton's method for the rest. In Chapter 2, we have seen that under favorable conditions, we may know when gradient descent has taken us close enough.

In practice, Newton's method is often (but not always) much faster than gradient descent in terms of the number of iterations. The price to pay is a higher iteration cost, since we need to compute (and invert) Hessians.

After this disclaimer, let us state the main result right away. We follow Vishnoi [Vis14]

**Theorem 6.4.** *Let $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$ be convex with a unique global minimum $\mathbf{x}^\star$. Suppose that there is an open ball $X \subseteq \mathbf{dom}(f)$ with center $\mathbf{x}^\star$ such that the following two properties hold.*

*(i)* Bounded inverse Hessians: *There exists a real number $\mu > 0$ such that*

$$\|\nabla^2 f(\mathbf{x})^{-1}\| \leq \frac{1}{\mu}, \quad \forall \mathbf{x} \in X.$$

*(ii)* Lipschitz continuous Hessians*: There exists a real number $L > 0$ such that*

$$\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

*In both cases, the matrix norm is the spectral norm defined in Lemma 2.4. Property (i) in particular stipulates that Hessians are invertible at all points in $X$.*

*Then, for $\mathbf{x}_t \in X$ and $\mathbf{x}_{t+1}$ resulting from the Newton step (6.6), we have*

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{L}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Before we prove this, here is the local convergence result that follows.

**Corollary 6.5** (Exercise 28)**.** *With the assumptions and terminology of Theorem 6.4, and if*

$$\|\mathbf{x}_0 - \mathbf{x}^\star\| < \frac{\mu}{L},$$

*then Newton's method (6.6) yields*

$$\|\mathbf{x}_T - \mathbf{x}^\star\| < \frac{2\mu}{L}\left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0.$$

Hence, we have a bound as (6.4) for the last phase of the Babylonian method: in order to get $\|\mathbf{x}_T - \mathbf{x}^\star\| < \varepsilon$, we only need $T = \log\log(\frac{1}{\varepsilon})$ steps. But before this fast behavior kicks in, we need to be $\mu/L$-close to $\mathbf{x}^\star$ already.

Towards the proof of Theorem 6.4, we need one more small tool.

**Lemma 6.6** (Exercise 29)**.** *Let $f$ be twice differentiable over a convex domain* $\mathbf{dom}(f)$*, $\mathbf{x}, \mathbf{y} \in$* $\mathbf{dom}(f)$*. Then*

$$\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})dt = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}). \tag{6.7}$$

*Proof of Theorem 6.4.* To simplify notation, let us abbreviate $H := \nabla^2 f$, $\mathbf{x} = \mathbf{x}_t, \mathbf{x}' = \mathbf{x}_{t+1}$. Subtracting $\mathbf{x}^\star$ from both sides of (6.6), we get

$$\begin{aligned} \mathbf{x}' - \mathbf{x}^\star &= \mathbf{x} - \mathbf{x}^\star - H(\mathbf{x})^{-1}\nabla f(\mathbf{x}) \\ &= \mathbf{x} - \mathbf{x}^\star + H(\mathbf{x})^{-1}(\nabla f(\mathbf{x}^\star) - \nabla f(\mathbf{x})) \\ &= \mathbf{x} - \mathbf{x}^\star + H(\mathbf{x})^{-1}\int_0^1 H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x}))(\mathbf{x}^\star - \mathbf{x})dt, \end{aligned}$$

using Lemma 6.6. With

$$\mathbf{x} - \mathbf{x}^\star = H(\mathbf{x})^{-1}H(\mathbf{x})(\mathbf{x} - \mathbf{x}^\star) = H(\mathbf{x})^{-1}\int_0^1 -H(\mathbf{x})(\mathbf{x}^\star - \mathbf{x})dt,$$

we further get

$$\mathbf{x}' - \mathbf{x}^\star = H(\mathbf{x})^{-1}\int_0^1 \left(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right)(\mathbf{x}^\star - \mathbf{x})dt.$$

Taking norms, we have

$$\|\mathbf{x}' - \mathbf{x}^\star\| \leq \|H(\mathbf{x})^{-1}\| \cdot \left\| \int_0^1 \left(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\right)(\mathbf{x}^\star - \mathbf{x})dt \right\|,$$

where we have used that $\|A\mathbf{y}\| \leq \|A\| \cdot \|\mathbf{y}\|$ for any matrix $A \in \mathbb{R}^{d\times d}$ and any vector $\mathbf{y} \in \mathbb{R}^d$ which follows directly from the definition of the spectral norm. As we also have

$$\left\| \int_0^1 \mathbf{g}(t)dt \right\| \leq \int_0^1 \|\mathbf{g}(t)\|dt$$

for any vector-valued function g (Exercise 32), we can further bound

$$\begin{aligned} \|\mathbf{x}' - \mathbf{x}^\star\| &\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \left\| \big(H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x})\big)(\mathbf{x}^\star - \mathbf{x}) \right\| dt \\ &\leq \|H(\mathbf{x})^{-1}\| \int_0^1 \left\| H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x}) \right\| \cdot \|\mathbf{x}^\star - \mathbf{x}\| dt \\ &\leq \|H(\mathbf{x})^{-1}\| \cdot \|\mathbf{x}^\star - \mathbf{x}\| \int_0^1 \left\| H(\mathbf{x} + t(\mathbf{x}^\star - \mathbf{x})) - H(\mathbf{x}) \right\| dt. \end{aligned}$$

We can now use the properties (i) and (ii) (bounded inverse Hessians, Lipschitz continuous Hessians) to conclude that

$$\|\mathbf{x}' - \mathbf{x}^\star\| \leq \frac{1}{\mu}\|\mathbf{x}^\star - \mathbf{x}\| \int_0^1 L\|t(\mathbf{x}^\star - \mathbf{x})\|dt = \frac{L}{\mu}\|\mathbf{x}^\star - \mathbf{x}\|^2 \underbrace{\int_0^1 t dt}_{1/2}.$$

□

How realistic are properties (i) and (ii)? If $f$ is twice *continuously* differentiable (meaning that the second derivative $\nabla^2 f$ is continuous), then we will always find suitable values of $\mu$ and $L$ over an open ball $X$ with center $\mathbf{x}^\star$—provided that $\nabla^2 f(\mathbf{x}^\star) \neq 0$.

Indeed, already in the one-dimensional case, we see that under $f''(x^\star) = 0$ (vanishing second derivative at the global minimum), Newton's method will in the worst reduce the distance to $x^\star$ at most by a constant factor in each step, no matter how close to $x^\star$ we start. Exercise 31 asks you to find such an example. In such a case, we have linear convergence, but the fast quadratic convergence ($\mathcal{O}(\log\log(1/\varepsilon))$ steps cannot be proven.

One way to ensure bounded inverse Hessians is to require strong convexity over $X$.

**Lemma 6.7** (Exercise 33)**.** *Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable and strongly convex with parameter $\mu$ over an open convex subset $X \subseteq \mathbf{dom}(f)$ according to Definition 3.5, meaning that*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

*Then $\nabla^2 f(\mathbf{x})$ is invertible and $\|\nabla^2 f(\mathbf{x})^{-1}\| \leq 1/\mu$ for all $\mathbf{x} \in X$, where $\|\cdot\|$ is the spectral norm defined in Lemma 2.4.*

## 6.4 Exercises

**Exercise 26.** *Consider the Babylonian method (6.2). Prove that we get $x_T - \sqrt{R} < 1/2$ for $T = \mathcal{O}(\log R)$.*

**Solution:** We subdivide the analyis into two phase. In the first phase, we are waiting for $x_t < \frac{3}{2}\sqrt{R}$. As long as $x_t \geq \frac{3}{2}\sqrt{R}$, we have

$$x_{t+1} = \frac{1}{2}\left(x_t + \frac{R}{x_t}\right) \leq \frac{1}{2}\left(x_t + \frac{2}{3}\sqrt{R}\right) \leq \frac{1}{2}\left(x_t + \frac{4}{9}x_t\right) = \frac{13}{18}x_t.$$

Hence, for $T = \mathcal{O}(\log R)$, we have $x_t < \frac{3}{2}\sqrt{R}$. For the analysis of the second phase, assume that $x_0 < \frac{3}{2}\sqrt{R}$. Dividing inequality (6.3) by $\sqrt{R}$ yields the relative error bound

$$\frac{x_{t+1} - \sqrt{R}}{\sqrt{R}} = \frac{\sqrt{R}}{2x_t}\left(\frac{x_t - \sqrt{R}}{\sqrt{R}}\right)^2 \leq \left(\frac{x_t - \sqrt{R}}{\sqrt{R}}\right)^2,$$

since we always have $x_t \geq \sqrt{R}$. Hence

$$\frac{x_T - \sqrt{R}}{\sqrt{R}} \leq \left(\frac{x_o - \sqrt{R}}{\sqrt{R}}\right)^{2^T} \leq \left(\frac{1}{2}\right)^{2^T}, \quad T \geq 0.$$

In order to have $x_T - \sqrt{R} < 1/2$, it suffices that

$$\left(\frac{1}{2}\right)^{2^T} < \frac{1}{2\sqrt{R}},$$

and this requires $T = \mathcal{O}(\log\log R)$ steps. In total, we therefore need $\mathcal{O}(\log R)$ steps, but the second phase is exponentially faster.

**Exercise 27.** *Prove Lemma 6.2!*

**Solution:** Using the chain rule of multivariate calculus, we compute

$$\begin{aligned}
\nabla(f \circ g)(\mathbf{y}) &= A^\top \nabla f(g(\mathbf{y})), \\
\nabla^2(f \circ g)(\mathbf{y}) &= A^\top \nabla^2 f(g(\mathbf{y}))A.
\end{aligned}$$

Hence,

$$\begin{aligned}
N_{f\circ g}(\mathbf{y}) &= \mathbf{y} - (A^\top \nabla^2 f(g(\mathbf{y}))A)^{-1} A^\top \nabla f(g(\mathbf{y})) \\
&= \mathbf{y} - A^{-1}\nabla^2 f(g(\mathbf{y}))^{-1}\nabla f(g(\mathbf{y})) \\
&= A^{-1}(g(\mathbf{y}) - \mathbf{b}) - A^{-1}\nabla^2 f(g(\mathbf{y}))^{-1}\nabla f(g(\mathbf{y})) \\
&= A^{-1}(g(\mathbf{y}) - \nabla^2 f(g(\mathbf{y}))^{-1}\nabla f(g(\mathbf{y}))) - A^{-1}\mathbf{b} \\
&= A^{-1}N_f(g(\mathbf{y})) - A^{-1}\mathbf{b} \\
&= A^{-1}(N_f(g(\mathbf{y})) - \mathbf{b}) \\
&= g^{-1}(N_f(g(\mathbf{y}))).
\end{aligned}$$

**Exercise 28.** *Prove Corollary 6.5!*

**Solution:**

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \frac{L}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

implies

$$\frac{L}{2\mu}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\| \leq \left(\frac{L}{2\mu}\|\mathbf{x}_t - \mathbf{x}^\star\|\right)^2.$$

Hence,

$$\frac{L}{2\mu}\|\mathbf{x}_T - \mathbf{x}^\star\| \leq \left(\frac{L}{2\mu}\|\mathbf{x}_0 - \mathbf{x}^\star\|\right)^{2^T} < \left(\frac{1}{2}\right)^{2^T},$$

or

$$\|\mathbf{x}_T - \mathbf{x}^\star\| < \frac{2\mu}{L}\left(\frac{1}{2}\right)^{2^T}.$$

**Exercise 29.** *Prove Lemma 6.6!* ***Hint:*** *You may use*

$$\int_a^b g'(t)dt = g(b) - g(a),$$

*where $g'$ is the derivative of the univariate function $g$.*

**Solution:** If $f : \mathbf{dom}(f) \to \mathbb{R}$ is a multivariate function and $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$ are fixed, we apply the hint componentwise to the vector-valued function $\mathbf{g}(t) = (g_1(t), \ldots, g_d(t)) := \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$. The chain rule of multivariate calculus yields

$$(g_1'(t), \ldots, g_d'(t)) = \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}),$$

hence the hint yields

$$\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x})dt = \mathbf{g}(1) - \mathbf{g}(0) = \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}).$$

**Exercise 30.** *Prove Lemma 6.3!*

**Solution:** Let $g(\mathbf{x}) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x} - \mathbf{x}_t) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t)$. We have

$$\nabla^2 g(\mathbf{x}) = \nabla^2 f(\mathbf{x}_t) \succ 0,$$

so $g$ is convex over $\mathbb{R}^d$ by Lemma 1.8. By Lemma 1.12, any solution to $\nabla g(\mathbf{x}) = \mathbf{0}$ is a global minimum of $g$. It remains to show that $\nabla g(\mathbf{x}_{t+1}) = \mathbf{0}$. We compute

$$\nabla g(\mathbf{x}) = \nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x} - \mathbf{x}_t),$$

hence

$$\begin{aligned}\nabla g(\mathbf{x}_{t+1}) &= \nabla f(\mathbf{x}_t) + \nabla^2 f(\mathbf{x}_t)(\mathbf{x}_{t+1} - \mathbf{x}_t) \\ &= \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)\nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t) = \mathbf{0}.\end{aligned}$$

**Exercise 31.** *Let $\delta > 0$ be any real number. Find an example of a convex function $f : \mathbb{R} \to \mathbb{R}$ such that (i) the unique global minimum $x^\star$ has a vanishing second derivative $f''(x^\star) = 0$, and (ii) Newton's method satisfies*

$$|x_{t+1} - x^\star| \geq (1-\delta)|x_t - x^\star|,$$

*for all $x_t \neq x^\star$.*

**Solution:** We take $f(x) = x^k$ for some even natural number $k$ satisfying $k \geq 4$ and $1/(k-1) \leq \delta$. We have

$$\begin{aligned}f'(x) &= kx^{k-1}, \\ f''(x) &= k(k-1)x^{k-2} \geq 0,\end{aligned}$$

hence $f$ is convex by the second-order characterization (1.8), and we have $x^\star = 0$ as well as $f''(x^\star) = 0$. Suppose w.l.o.g. that $x_t > 0$. The Newton step (6.1) is

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - \frac{kx_t^{k-1}}{k(k-1)x_t^{k-2}} = x_t - \frac{1}{k-1}x_t \geq (1-\delta)x_t.$$

**Exercise 32.** *This exercise is just meant to recall some basics around integrals. Show that for a vector-valued function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^d$, the inequality*

$$\left\|\int_0^1 \mathbf{g}(t)dt\right\| \leq \int_0^1 \|\mathbf{g}(t)\|dt$$

*holds, where $\|\cdot\|$ is the 2-norm (always assuming that the funtions under consideration are integrable)! You may assume (i) that integrals are linear:*

$$\int_0^1 (\lambda_1 g_1(t) + \lambda_2 g_2(t))dt = \lambda_1 \int_0^1 g_1(t)dt + \lambda_2 \int_0^1 g_2(t)dt,$$

*And (ii), if $g(t) \geq 0$ for all $t \in [0,1]$, then $\int_0^1 g(t)dt \geq 0$.*

**Solution:** Let $\mathbf{v} = \int_0^1 \mathbf{g}(t)dt$. If $\mathbf{v} = \mathbf{0}$, the statement follows from (i). Otherwise, we have

$$\begin{aligned}
\|\mathbf{v}\|^2 &= \mathbf{v}^T\mathbf{v} \\
&= \mathbf{v}^T \int_0^1 \mathbf{g}(t)dt \\
&= \int_0^1 \mathbf{v}^T\mathbf{g}(t)dt \quad \text{((i))} \\
&\leq \int_0^1 \|\mathbf{v}\| \cdot \|\mathbf{g}(t)\|dt \quad \text{(Cauchy-Schwarz inequality, (i), (ii))} \\
&= \|\mathbf{v}\| \int_0^1 \|\mathbf{g}(t)\|dt \quad \text{((i))}.
\end{aligned}$$

Dividing by $\|\mathbf{v}\|$, the statement follows.

**Exercise 33.** *Prove Lemma 6.7! You may want to proceed in the following steps.*

*(i) Prove that the function $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex over $X$.*

*(ii) Prove that $\nabla^2 f(\mathbf{x})$ is invertible for all $\mathbf{x} \in X$.*

*(iii) Prove that all eigenvalues of $\nabla^2 f(\mathbf{x})^{-1}$ are positive and at most $1/\mu$.*

*(iv) Prove that for a symmetric matrix $M$, the spectral norm $\|M\|$ is the largest absolute eigenvalue.*

**Solution:** (i) For all $\mathbf{x}, \mathbf{y} \in X$, we use strong convexity to derive

$$\begin{aligned}
g(\mathbf{y}) &= f(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|^2 \\
&\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 - \frac{\mu}{2}\|\mathbf{y}\|^2 \\
&= g(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{x}\|^2 + (\nabla g(\mathbf{x}) + \mu\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 - \frac{\mu}{2}\|\mathbf{y}\|^2 \\
&= g(\mathbf{x}) + \nabla g(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2 + \mu\mathbf{x}^\top\mathbf{y} + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2 - \frac{\mu}{2}\|\mathbf{y}\|^2 \\
&= g(\mathbf{x}) + \nabla g(\mathbf{x})^\top(\mathbf{y} - \mathbf{x}),
\end{aligned}$$

by the equation $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$. Hence, $g$ is convex over $X$ by the first order characterization of convexity (Lemma 1.7).

(ii) With $g$ as in (i), we have $\nabla^2 f(\mathbf{x}) = \nabla^2 g(\mathbf{x}) + \mu I$, where $I$ is the identity matrix. Since $g$ is convex, $\nabla^2 g(\mathbf{x})$ is positive semidefinite by the second-order characterization of convexity (Lemma 1.8). Then, $\nabla^2 f(\mathbf{x})$ is positive definite, since

$$\mathbf{y}^\top \nabla^2 f(\mathbf{x})\mathbf{y} = \underbrace{\mathbf{y}^\top \nabla^2 g(\mathbf{x})\mathbf{y}}_{\geq 0} + \mu\|\mathbf{y}\|^2 > 0, \quad \forall \mathbf{y} \neq \mathbf{0}.$$

A positive definite matrix $M$ is invertible, since there can't be a nonzero $\mathbf{y}$ satisfying $M\mathbf{y} = \mathbf{0}$.

(iii) We first show that all eigenvalues of $\nabla^2 f(\mathbf{x})$ are at least $\mu$. Let $\lambda$ be an eigenvalue with nonzero eigenvector $\mathbf{v}$. Then we get

$$\lambda\|\mathbf{v}\|^2 = \mathbf{v}^T \nabla^2 f(\mathbf{x})\mathbf{v} = \underbrace{\mathbf{v}^T \nabla^2 g(\mathbf{x})\mathbf{v}}_{\geq 0} + \mu\|\mathbf{v}\|^2,$$

hence $\lambda \geq \mu$. Furthermore, $\lambda$ is an eigenvalue of an invertible matrix $M$ if and only if $1/\lambda$ is an eigenvalue of $M^{-1}$. Hence, all eigenvalues of $\nabla^2 f(\mathbf{x})^{-1}$ are at most $1/\mu$.

(iv) We have

$$\|M\| = \max_{\|\mathbf{x}\|=1} \|M\mathbf{x}\|,$$

hence

$$\|M\|^2 = \max_{\|\mathbf{x}\|=1} \|M\mathbf{x}\|^2 = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top M^\top M\mathbf{x}.$$

From linear algebra we know that a symmetric matrix $M$ can be diagonalized: there exists an orthonormal matrix $U$ (orthonormal means that $U^{-1} = U^\top$), and a diagonal matrix $D$ with the eigenvalues of $M$ on the diagonal, such that $M = U^\top DU$. As an orthonormal matrix preserves norms ($\|U\mathbf{x}\|^2 = \mathbf{x}^\top U^\top U\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|x\|^2$), we further get

$$\begin{aligned}
\|M\|^2 &= \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top M^\top M\mathbf{x} \\
&= \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top U^T D^2 U\mathbf{x} \\
&= \max_{\|U\mathbf{x}\|=1} \mathbf{x}^\top U^T D^2 U\mathbf{x} \\
&= \max_{\|\mathbf{y}\|=1} \mathbf{y}^\top D^2 \mathbf{y} \\
&= \max_{\|\mathbf{y}\|=1} \sum_{i=1}^{d} y_i^2 \lambda_i^2,
\end{aligned}$$

where the $\lambda_i$ are the eigenvalues of $M$. The sum is maximized when $y_i = 1$ for the largest $\lambda_i^2$, hence

$$\|M\|^2 = \max_{i=1}^{d} \lambda_i^2, \quad \|M\| = \max_{i=1}^{d} |\lambda_i|.$$

# Bibliography

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. https://web.stanford.edu/~boyd/cvxbook/.

[Dav59] William C. Davidon. Variable metric method for minimization. Technical Report ANL-5990, AEC Research and Development, 1959.

[Dav91] William C. Davidon. Variable metric method for minimization. *SIAM J. Optimization*, 1(1):1–17, 1991.

[DSSSC08] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the 1-ball for learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning*, pages 272–279, 07 2008.

[Gol70] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

[Gre70] J. Greenstadt. Variations on variable-metric methods. *Mathematics of Computation*, 24(109):1–22, 1970.

[KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear Convergence of Gradient and Proximal-Gradient Methods Under the Polyak-Łojasiewicz Condition. In *ECML PKDD 2016: Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer International Publishing, Cham, September 2016.

[Nes12] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[Noc80] J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[NP06] Yurii Nesterov and B.T. Polyak. Cubic regularization of newton method and its global performance. *Mathematical Programming*, 108(1):177–205, Aug 2006.

[NSL$^+$15] Julie Nutini, Mark W Schmidt, Issam H Laradji, Michael P Friedlander, and Hoyt A Koepke. Coordinate Descent Converges Faster with the Gauss-Southwell Rule Than Random Selection. In *ICML*, pages 1632–1641, 2015.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[Vis14] Nisheeth Vishnoi. Lecture notes on fundamentals of convex optimization, 2014. https://tcs.epfl.ch/files/content/sites/tcs/files/Lec3-Fall14-Web.pdf.

[Zim16] Judith Zimmermann. *Information Processing for Effective and Stable Admission*. PhD thesis, ETH Zurich, 2016. .