Lecture 12. Newton's Method

12.1 Newton's method for optimization

Recall that the gradient descent converges slowly if the condition number is large. For example, consider the function $f(x_1, x_2) = \frac{1}{100}x_1^2 + x_2^2$. At some point (y_1, y_2), $-\nabla f(y_1, y_2) = \left(-\frac{1}{50}y_1, -2y_2\right)$. It locally decreases rapidly but not globally. The ideal descending direction is $\boldsymbol{d} = (-y_1, -y_2)$, which can also be written as

$$\boldsymbol{d} = -\begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \nabla f(y_1, y_2) = -\left(\nabla^2 f(y_1, y_2)\right)^{-1} \nabla f(y_1, y_2)\,.$$

In general, if $f(x) = x^\top Q x$ where $Q \succ 0$, the ideal direction at $\boldsymbol{x} = (x_1, x_2)$ is

$$\boldsymbol{d} = -(x_1, x_2) = -\frac{1}{2}Q^{-1}\nabla f(x)\,,$$

since $\nabla f(x) = 2Qx$.

More generally, recall the Newton's method introduced before for finding roots, where we use a Taylor series to estimate the objective function. When $f(x)$ is not quadratic, consider its Taylor approximation

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_k) + \nabla f(\boldsymbol{x}_k)^\top(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^\top \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k)\,.$$

Denote by $g(\boldsymbol{x})$ the right hand side. Note that

$$\begin{aligned}\nabla g(\boldsymbol{x}) = 0 &\iff \nabla f(\boldsymbol{x}_k) + \nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x} - \boldsymbol{x}_k) = 0 \\ &\iff \boldsymbol{x} = \boldsymbol{x}_k - \left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k)\end{aligned}$$

if $\nabla^2 f(\boldsymbol{x}_k)$ is invertible. Thus the minimizer of $g(\boldsymbol{x})$ is $\boldsymbol{x} = \boldsymbol{x}_k - \left(\nabla^2 f(x_k)\right)^{-1}\nabla f(x_k)$. Then, we let $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \left(\nabla^2 f(x_k)\right)^{-1}\nabla f(\boldsymbol{x}_k)$ be the minimizer of the second-order Taylor series at $\boldsymbol{x}_k$. This method is called the *Newton's method* for optimization.

Note that if the objective function is *strictly convex*, then $\nabla^2 f(x_k) \succ 0$, which implies that $\nabla^2 f(x_k)$ is invertible and $\left(\nabla^2 f(x_k)\right)^{-1} \succ 0$. This is because if $\nabla^2 f(x_k)$'s eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are all positive, their inverse $\lambda_1^{-1}, \lambda_2^{-1}, \ldots, \lambda_n^{-1}$ are also positive. Therefore, $\left(\nabla^2 f(x_k)\right)^{-1} \succ 0$.

Recall our requirement for the descending direction. We hope $\langle \nabla^2 f(x_k), \boldsymbol{d} \rangle < 0$. Clearly, if $\left(\nabla f(x_k)\right)^{-1} \succ 0$, $\boldsymbol{d} = -\left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)$ is a descending direction, since

$$\langle \nabla f(x_k), \boldsymbol{d} \rangle = -\nabla f(x_k)^\mathsf{T} \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k) < 0\,,$$

unless $\nabla f(x_k) = \mathbf{0}$.

12.2 Convergence rate of the Newton's method

Question

Does the Newton's method always work well?

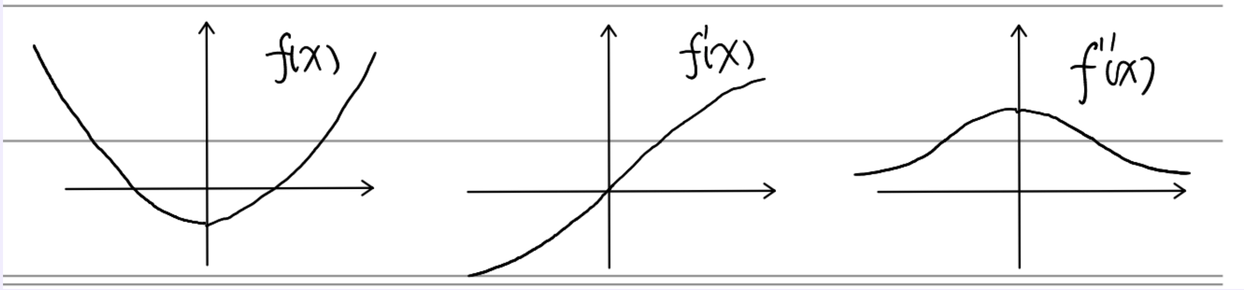
Intuitively this is not true since we use the second order Taylor series to approximate the function, but the Taylor series only works locally.

If the second-order Taylor series estimates the value of functions well, then x_{k+1} given by the Newton's method is the minimum point of the Taylor series and it should be close to the minimum point of f. But what happens if the Taylor series doesn't approximates well? Now we consider some "bad" examples.

Example

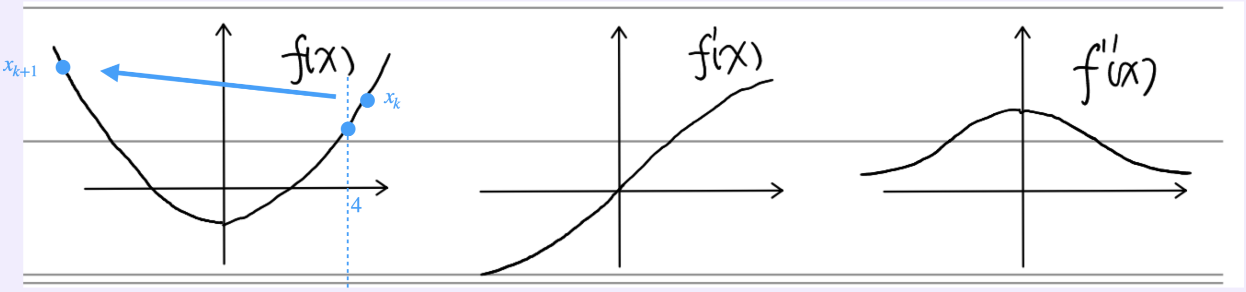
Consider the function $f(x) = x \sinh^{-1} - \sqrt{1 + x^2} = x \ln\left(x + \sqrt{x^2 + 1}\right) - \sqrt{x^2 + 1}$.

- Its first-order derivative is $f'(x) = \sinh^{-1}(x)$;
- Its second-order derivative is $f''(x) = \frac{1}{x^2+1}$.



If we set $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$, then we can show that $|x_{k+1}| > |x_k|$ as long as

$|x_k| \geq 4$. Thus, the Newton's method does not converge in this example.



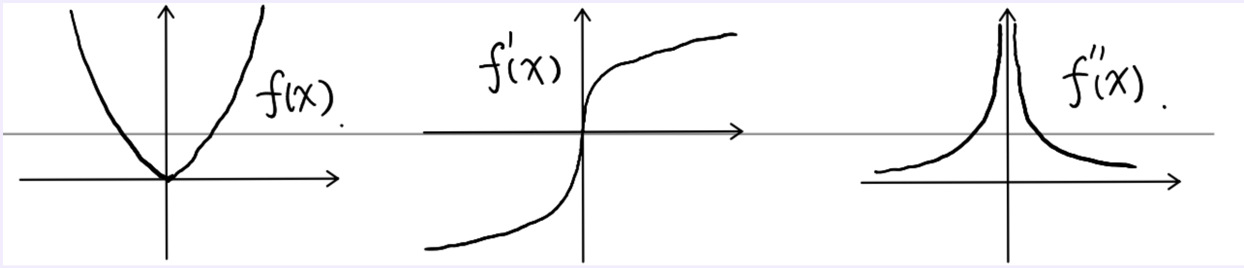
The reason why the Newton's method does not converge in this example is that the second-order derivative of f is close to zero when $|x|$ is large, which yields that the second-order Taylor series is not a good approximation near the minimum point. The Taylor series approximates well in the neighborhood of x_k, but loses control at x^*, if $|x_k - x^*|$ is large.

However, just keeping $|x_k - x^*|$ small is not enough yet. Here is another "bad" example.

Example

Consider the function $f(x) = x^{\frac{4}{3}}$.

- Its first-order derivative is $f'(x) = \frac{4}{3}x^{\frac{1}{3}}$;
- Its second-order derivative is $f''(x) = \frac{4}{9}x^{-\frac{2}{3}}$.



In this case, $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = -2x_k$. Clearly, the Newton's method does not produce convergent iterates.

In this example, the reason to failure is that $f''(x)$ changes rapidly so the second order Taylor series cannot approximates $f(x)$ well even in the neighborhood of x_k.

We now give some conditions to guarantee the convergence of the Newton's method iterates.

Definition

Given a twice continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, we say $\nabla^2 f(x)$ is *M-Lipschitz*, if for all $x, y \in \mathbb{R}^n$,

$$\|\nabla^2 f(x) - \nabla^2 f(y)\| \leq M \cdot \|x - y\|_2 \,.$$

Theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a μ-strongly convex function, and $\nabla^2 f$ is M-Lipschitz. Let $\{x_k\}$ be the iterates generated by the Newton's method. Then

$$\|x_{k+1} - x^*\| \leq \frac{M}{2\mu} \|x_k - x^*\|^2 \,.$$

Remark

Let $y_k = \frac{M}{2\mu}\|x_k - x^*\|$. Then we have $y_{k+1} \leq y_k^2 \leq y_0^{2^{k+1}}$. So if $y_0 < 1$, the iterates given by the Newton's method converge rapidly (much more rapidly than what we get for the gradient descent).

This is called the *quadratic convergence*, or the *convergence of order* 2.

Proof

Fix k. Let $g(t) = \nabla f(x^* + t(x_k - x^*))$. Then $g'(t) = \left(\nabla^2 f(x^* + t(x_k - x^*))\right)(x_k - x^*)$. So applying the Newton–Leibniz formula, we have

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k - x^* - \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k)\| \\
&= \|\left(\nabla^2 f(x_k)\right)^{-1} \left(\nabla^2 f(x_k)(x_k - x^*)\right) - \left(\nabla^2 f(x_k)\right)^{-1} (\nabla f(x_k) - \nabla f \\
&\leq \|\nabla^2 f(x_k)^{-1}\| \cdot \|\nabla^2 f(x_k)(x_k - x^*) - (g(1) - g(0))\| \\
&= \|\nabla^2 f(x_k)^{-1}\| \cdot \|g'(1) - \int_0^1 g'(t)\,\mathrm{d}t\| \\
&\leq \frac{1}{\mu} \cdot \|\int_0^1 (Dg(1) - Dg(t))\,\mathrm{d}t\| \\
&\leq \frac{1}{\mu} \cdot \int_0^1 \|g'(1) - g'(t)\|\,\mathrm{d}t \\
&= \frac{1}{\mu} \cdot \int_0^1 \|\left(\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\right) \cdot (x_k - x^*)\|\,\mathrm{d}t \\
&\leq \frac{1}{\mu} \cdot \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\| \cdot \|x_k - x^*\|\,\mathrm{d}t \\
&\leq \frac{1}{\mu} \int_0^1 M(1-t)\|x_k - x^*\| \cdot \|x_k - x^*\|\,\mathrm{d}t \\
&= \frac{M}{2\mu} \cdot \|x_k - x^*\|^2 .
\end{aligned}$$

Recall the gradient descent iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)$, where we actually calculate the minimum point of the following function

$$\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x} - \boldsymbol{x}_k \rangle + \frac{1}{2\eta}\|\boldsymbol{x} - \boldsymbol{x}_k\|^2 ,$$

and let $\boldsymbol{x}_{k+1} = \arg\min_{\boldsymbol{x}} \hat{f}(\boldsymbol{x})$ be its minimizer to approximate the minimum point of $f(\boldsymbol{x})$. In particular, if $\eta \leq 1/L$, $\hat{f}(\boldsymbol{x})$ is an upper estimation of $f(\boldsymbol{x})$ by L-smoothness.
The second order Taylor series locally approximate $f(\boldsymbol{x})$ well, if f satisfies some conditions, so the Newton's method converges rapidly near $\boldsymbol{x}^*$. However, if $\boldsymbol{x}_0$ is far from $\boldsymbol{x}^*$, the Newton's method loses control of $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|$. In contrast, L-smoothness guarantees a global upper bound so gradient descent iterations (with a fixed step size $\eta \leq 1/L$) always converge, although it does not converge as fast as Newton's method in the neighbourhood of $\boldsymbol{x}^*$.

Norm of matrices

Note that $\nabla^2 f$ is a $n \times n$ matrix if f is a function mapping $\mathbb{R}^n$ to $\mathbb{R}$. To describe the "rapid change" of $\nabla^2 f$, we need to define a norm of matrices.

A simple idea is to view a $n \times n$ matrix as a n^2-dimensional vector, and applying L^p -norms. If $p = 2$, such a norm is called the *Frobenius norm*.

Definition (*Frobenius norm*)

The *Frobenius norm*, sometimes also called the *Euclidean norm*, is the matrix norm of an $m \times n$ matrix A defined as the square root of the sum of the absolute squares of its elements, namely,

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \right)^{1/2} .$$

However, a more natural way is to consider the following definition. We may view an $m \times n$ matrix as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. So we can define its *operator norm* as follows.

Definition

Given a norm $\|\cdot\|_a$ on $\mathbb{R}^n$ and a norm $\|\cdot\|_b$ on $\mathbb{R}^m$, the *operator norm* of an $m \times n$ matrix M is given by

$$\|M\|_{a,b} = \max_{\boldsymbol{v} \neq \boldsymbol{0}} \frac{\|M\boldsymbol{v}\|_b}{\|\boldsymbol{v}\|_a} = \max_{\|\boldsymbol{v}\|_a = 1} \|M\boldsymbol{v}\|_b \, .$$

In particular, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are both L^2-norms, the operator norm $\|M\|_{a,b}$ is also called the *spectral norm*.

Unless specified in context, we use $\|M\|$ to denote the spectral norm. Why do we call it "spectral norm"?

Proposition

If we use $\lambda_{\max}(M)$ to denote the maximum eigenvalue of M, then

$$\|Q\| = \sqrt{\lambda_{\max}(Q^\top Q)} \, .$$

In particular, if $Q \succeq 0$, then $\|Q\| = \lambda_{\max}(Q)$.

The *spectrum* of a matrix is the set of all its eigenvalues. This proposition shows that why this norm is called the "spectral" norm.

Proof

We have $\|Q\boldsymbol{v}\|^2 = \langle Q\boldsymbol{v}, Q\boldsymbol{v}\rangle = (Q\boldsymbol{v})^\top Q\boldsymbol{v} = \boldsymbol{v}^\top Q^\top Q\boldsymbol{v} \leq \lambda_{\max}(Q^\top Q)\|\boldsymbol{v}\|^2$, since $Q^\top Q \succeq 0$.

The advantage to use operator norm is that we usually need the Cauchy-Schwarz inequality, which is trivially true (by definition) under the operator norm: for all $v \in \mathbb{R}^n$ and $Q \in \mathbb{R}^{m\times n}$, it holds that

$$\|Qv\| \leq \|Q\| \cdot \|v\|\,.$$

12.3 Damped Newton's method

Unfortunately, Newton's method does not guarantee descent of the function values even when the Hessian matrix is positive definite. Similar to the gradient descent with a step size η, we can modify the Newton's method to include a small step size $\eta \in (0, 1)$ instead of $\eta = 1$, where the step size η is chosen by a certain line search. This is called the *damped Newton's method*.

Since $\boldsymbol{d} = -\nabla^2 f(\boldsymbol{x}_k)\nabla f(\boldsymbol{x}_k)$ is a descending direction (by convexity), we claim that there exists $\eta > 0$ such that

$$f(\boldsymbol{x}_k + \eta\,\boldsymbol{d}) < f(\boldsymbol{x}) + \alpha\eta\,\langle\nabla f(\boldsymbol{x}_k),\, d\rangle$$

with parameter $\alpha \in (0, 1)$. Again, applying the backtracking line search, we can find such η by starting from an initial $\eta > 0$ (usually $\eta = 1$) and repeating $\eta \leftarrow \beta\eta$ until the above sufficient decrease condition is satisfied.

Convergence analysis

The convergence of the damped Newton's method has two phase: damped Newton phase and quadratically convergent phase. We can show that there exists $\delta \in \left(0, \frac{\mu^2}{M}\right)$ and $\gamma > 0$ such that the following holds. Specifically, assuming $0 < \delta < \min\left\{\frac{\mu^2}{M}, 3(1-2\alpha)\frac{\mu^2}{M}\right\}$ and $\gamma = \alpha\beta^2\delta^2\frac{\mu}{L^2}$ for a constant L satisfying $\nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}$ for any $\boldsymbol{x} \in \mathbb{R}^n$, we have

- (*damped Newton phase*) if $\|\nabla f(\boldsymbol{x}_k)\| \geq \delta$, then

$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \leq -\gamma\,;$$

- (*quadratically convergent phase*) if $\|\nabla f(\boldsymbol{x}_k)\| < \delta$, then the backtracking line search condition is satisfied by selecting $\eta = 1$, and

$$\|\nabla f(\boldsymbol{x}_{k+1})\| \leq \frac{M^2}{2\mu}\|\nabla f(\boldsymbol{x}_k)\|^2\,.$$

12.4 Self-concordant functions

Another way to control $\|\nabla^2 f(x) - \nabla^2 f(y)\|$ is to compute the third derivative. For simplicity, we consider a univariate function $f : \mathbb{R} \to \mathbb{R}$. If $|f'''(x)|$ is bounded then $f''(x)$ is Lipschitz. Previous analysis involves the bound of $f''(x)$ and $f'''(x)$ separately. We now introduce another assumption of functions, in which we take into consideration both $f'''(x)$ and $f''(x)$ simultaneously.

Moreover, Newton's method is *affinely invariant*. Suppose $\boldsymbol{T} \in \mathbb{R}^{n\times n}$ is nonsingular, and define $\bar{f}(\boldsymbol{y}) = f(\boldsymbol{T}\boldsymbol{y})$. If we use Newton's method (with the same backtracking parameters) to minimize $\bar{f}$, starting from $\boldsymbol{y}_0 = \boldsymbol{T}^{-1}\boldsymbol{x}_0$, then we have $\boldsymbol{T}\boldsymbol{y}_k = \boldsymbol{x}_k$ for all k. However, the previous convergence analysis is not affinely independent. In contrast, the following assumption does not depend on affine changes of coordinates.

Definition (*Self-concordant function*)

A convex function $f : \mathbb{R} \to \mathbb{R}$ is *self-concordant*, if

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

or, equivalently, f satisfies

$$\left|\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\sqrt{f''(x)}}\right| \leq 1$$

wherever $f''(x) > 0$ and satisfies $f'''(0) = 0$ elsewhere.
More generally, a multivariate convex function $f : \mathbb{R}^n \to \mathbb{R}$ is *self-concordant*, if it is self-concordant along every line in its domain, i.e., the function $g(t) = f(\boldsymbol{x} + t\boldsymbol{v})$ is a self-concordant function of t for all $\boldsymbol{x}$ and $\boldsymbol{v}$. Equivalently,

f is self-concordant, if

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla^2 f(\boldsymbol{x}+t\boldsymbol{v})\Big|_{t=0} = \lim_{s\to 0}\frac{1}{s}\left(\nabla^2 f(\boldsymbol{x}+s\boldsymbol{v}) - \nabla^2 f(\boldsymbol{x})\right) \preceq 2\sqrt{\boldsymbol{v}^\top \nabla^2 f(\boldsymbol{x})\boldsymbol{v}}\,\nabla^2 f(\boldsymbol{x})$$

The self-concordant functions include many of the logarithmic barrier functions that play an important role in barrier method and interior point method for solving convex optimization problems.

In fact, the coefficient 2 in the definition is not necessary, and it can be replaced by any constant $\kappa > 0$. The standard choice $\kappa = 2$ is to guarantee that $-\log x$ is a self-concordant function.

Example

- $f(x) = -\log x$ is self-concordant, since $f''(x) = \frac{1}{x^2}$ and $f'''(x) = -\frac{2}{x^3}$.
- $f(x) = x\log x - \log x$ is self-concordant.
- (*log-barrier for linear inequalities*) $f(\boldsymbol{x}) = -\sum_{i=1}^m \log(b_i - \boldsymbol{a}_i^\top \boldsymbol{x})$ on $\{\boldsymbol{x} \mid \boldsymbol{a}_i^\top \boldsymbol{x} < b_i, i = 1, 2, \ldots, m\}$ is self-concordant.
- (*log-determinant*) $f(\boldsymbol{X}) = -\log\det \boldsymbol{X}$ on $\mathcal{S}_{++}^n$ is self-concordant.
- $f(\boldsymbol{x}, y) = -\log(y^2 - \boldsymbol{x}^\top \boldsymbol{x})$ on $\{(\boldsymbol{x}, y) \mid \|\boldsymbol{x}\| \le y\}$ is self-concordant.
- If $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are both self-concordant, then $f + g$ is also self-concordant.

Convergence analysis

For strictly convex self-concordant function, we obtain bounds in terms of the Newton decrement

$$\lambda(\boldsymbol{x}) = \sqrt{\nabla f(\boldsymbol{x})^\top \left(\nabla^2 f(\boldsymbol{x})\right)^{-1} \nabla f(\boldsymbol{x})}\,.$$

There exists constants $\delta \in (0, 1/4]$ and $\gamma > 0$ (only depending on the backtracking line search parameters α and β) such that the following holds.

- (*damped Newton phase*) If $\lambda(\boldsymbol{x}_k) \ge \delta$, then
$$f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k) \le -\gamma\,;$$
- (*quadratically convergent phase*) If $\lambda(\boldsymbol{x}_k) < \delta$, then the backtracking line search condition is satisfied by selecting $\eta = 1$, and

$$\lambda(\boldsymbol{x}_{k+1}) \leq 2\lambda(\boldsymbol{x}_k)^2 \,.$$