Lecture 10. Convergence Rate of Gradient Descent

10.1 Gradient flow

At the end of the last lecture, we obtained the following lemma on the gradient descent for smooth functions.

Lemma (*Descent lemma*)

For an L-smooth differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ (not necessarily convex), and $\eta \leq 1/L$, we have

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

However, it is still not easy to show the convergence rate for the gradient descent. We now introduce a continuous version of the gradient descent instead, which is easier to analyse.

Definition (*Gradient flow*)

A *gradient flow* is a curve $\boldsymbol{x} : \mathbb{R} \to \mathbb{R}^n$ following the direction of steepest descent of a function. Given a smooth convex function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $\hat{\boldsymbol{x}} \in \mathbb{R}^n$, the *gradient flow* of f with initial point $\hat{\boldsymbol{x}}$ is the solution to the following differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{x}_t = -\nabla f(\boldsymbol{x}_t) , \qquad \boldsymbol{x}_0 = \hat{\boldsymbol{x}} .$$

Here we use the notation $\boldsymbol{x}_t = \boldsymbol{x}(t)$ for convenience.

Applying the chain rule, $f(\boldsymbol{x}_t)$ is decreasing since

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}_t) = \left\langle \nabla f(\boldsymbol{x}_t) , \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{x}_t \right\rangle = -\langle \nabla f(\boldsymbol{x}_t), \nabla f(\boldsymbol{x}_t) \rangle \leq 0 .$$

Now we can take the derivative

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|^2 &= 2\left\langle \boldsymbol{x}_t-\boldsymbol{x}^*, \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t\right\rangle \\ &= -2\left\langle \boldsymbol{x}_t-\boldsymbol{x}^*, \nabla f(\boldsymbol{x}_t)\right\rangle \\ &= 2\left\langle \boldsymbol{x}^*-\boldsymbol{x}_t, \nabla f(\boldsymbol{x}_t)\right\rangle \\ &\le 2\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_t)\big)\end{aligned}$$

by convexity. Then integrating both sides, we obtain that

$$\|\boldsymbol{x}_T-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2 \le 2\,T f(\boldsymbol{x}^*)-2\int_0^T f(\boldsymbol{x}_t)\,\mathrm{d}t \le 2\,T\big(f(\boldsymbol{x}^*)-f(\boldsymbol{x}_T)\big)\,,$$

which further gives that

$$f(\boldsymbol{x}_T)-f(\boldsymbol{x}^*) \le \frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2-\|\boldsymbol{x}_T-\boldsymbol{x}^*\|^2}{2\,T} \le \frac{\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2}{2\,T}\,.$$

10.2 Convergence of gradient descent with smoothness

We compare the gradient descent with the gradient flow. Assume the gradient descent iterates with a fixed step size η and an initial point $\hat{\boldsymbol{x}}$, i.e.,

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\eta\,\nabla f(\boldsymbol{x}_k)\,,\qquad \boldsymbol{x}_0=\hat{\boldsymbol{x}}\,.$$

For the gradient descent,

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-\int_{k\eta}^{(k+1)\eta}\nabla f(\boldsymbol{x}_k)\,\mathrm{d}t\,.$$

For the gradient flow,

$$\boldsymbol{x}_{(k+1)\eta}=\boldsymbol{x}_{k\eta}-\int_{k\eta}^{(k+1)\eta}\nabla f(\boldsymbol{x}_t)\,\mathrm{d}t\,.$$

Intuitively we know that, if ∇f does not change too fast, the gradient descent approximates the gradient flow.

Theorem

Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is a convex and L-smooth function. Choose $\eta\le 1/L$, and let the gradient descent iterate with a fixed step size η. Then it holds that

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2T\eta}.$$

Proof

Analogously to the gradient flow, we calculate $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2$. Note that

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 &= \langle \boldsymbol{x}_{k+1} - \boldsymbol{x}^*, \boldsymbol{x}_{k+1} - \boldsymbol{x}^* \rangle \\ &= \langle \boldsymbol{x}_k - \boldsymbol{x}^* - \eta \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* - \eta \nabla f(\boldsymbol{x}_k) \rangle \\ &= \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 - 2\eta \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}_k) \rangle + \eta^2 \|\nabla f(\boldsymbol{x}_k)\|^2.\end{aligned}$$

Similarly, it suffices to bound $-2\eta \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}_k) \rangle + \eta^2 \|\nabla f(\boldsymbol{x}_k)\|^2$ by $f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*)$. Since $\eta \leq 1/L$, we have

$$\eta^2 \|\nabla f(\boldsymbol{x}_k)\|^2 \leq 2\eta \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \right)$$

by the descent lemma. In addition, the convexity of f gives that

$$-2\eta \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}_k) \rangle = 2\eta \langle \boldsymbol{x}^* - \boldsymbol{x}_k, \nabla f(\boldsymbol{x}_k) \rangle \leq 2\eta \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \right).$$

Thus, we obtain that

$$\begin{aligned}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 &= -2\eta \langle \boldsymbol{x}_k - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}_k) \rangle + \eta^2 \|\nabla f(\boldsymbol{x}_k)\|^2 \\ &\leq 2\eta \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \right) + 2\eta \left(f(\boldsymbol{x}_k) - f(\boldsymbol{x}_{k+1}) \right) \\ &= 2\eta \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_{k+1}) \right).\end{aligned}$$

Summing over both sides from 0 to $T-1$, it implies that

$$\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \leq \sum_{k=1}^{T} 2\eta \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k) \right) \leq 2T\eta \left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_T) \right),$$

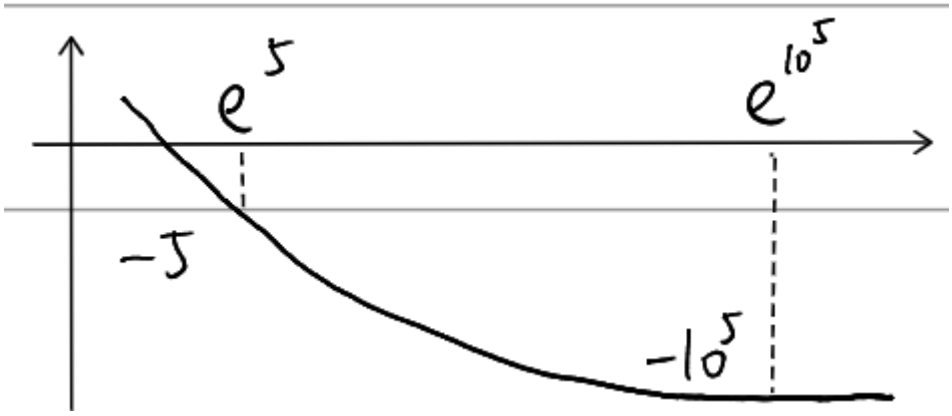
which is equivalent to

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2}{2T\eta}.$$

If we hope $|f(\boldsymbol{x}_T) - f^*| < \varepsilon$, we need to run the gradient descent $T = \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2}{2\varepsilon\eta} = O(1/\varepsilon)$ steps. If the initial point $\boldsymbol{x}_0$ is far from $\boldsymbol{x}^*$, and ε is sufficiently small, the gradient descent is slow. Unfortunately, consider the

following function

$$f(x) = \begin{cases} -\log x\,, & x < e^{10^5} \\ -10^5\,, & x \geqslant e^{10^5}\,. \end{cases}$$



This function is convex and 1-smooth. Hence $x_{k+1} = x_k + \eta/x_k \leq x_k + 1/x_k$ and the convergence rate of the gradient descent will be very small if $x_0 = 1$.

Question

Under which assumptions the gradient descent converges rapidly?

10.3 Strongly convex functions

Recall that, if we run the gradient descent for a quadratic function $f(x) = ax^2$ where $a \in \mathbb{R}_{>0}$, it gives that $x_{k+1} = (1 - 2a\eta)x_k$ and thus $f(x_k) = a(1 - 2a\eta)^{2k}x_0^2$. Clearly $f(x_k)$ converges to the optimal value 0 at an exponential rate.

We now introduce the following definition, which requires the function is a bit "better" than some quadratic function.

Definition (*Strong convexity*)

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex* with $\mu > 0$ if $f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

There are some other forms of quadratic functions. Why don't we choose other functions such as $x^\mathsf{T}Qx$ or $\|x - y_0\|^2$ for some given y_0? In fact, these functions mentioned can just achieve a similar effect to $\frac{\mu}{2}\|x\|^2$. For example, $x^\mathsf{T}Qx$ is almost equivalent to $\lambda_{\max}(Q)x^\mathsf{T}x = \lambda_{\max}(Q)\|x\|^2$. In addition,

$$\|x-y_0\|^2 = \|x\|^2 + \underbrace{\|y_0\|^2}_{\text{constant}} - \underbrace{2\langle x, y_0\rangle}_{\text{don't affect convexity}} .$$

Hence, all quadratic functions achieve similar effects to $\|x\|$.

Recall that, a function is convex iff its hessian matrix is positive semidefinite. The hessian matrix of $f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|^2$ is

$$\nabla^2 f(x) - \frac{\mu}{2}\nabla^2\left(\|\boldsymbol{x}\|^2\right) = \nabla^2 f(\boldsymbol{x}) - \frac{\mu}{2}\nabla^2\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}\right) = \nabla^2 f(\boldsymbol{x}) - \mu I .$$

Lemma

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function. Then f is μ-strongly convex iff $\nabla^2 f(\boldsymbol{x}) \succeq \mu \boldsymbol{I}_n$. Namely, for all $\boldsymbol{x} \in \mathbb{R}^n$, $\lambda_{\min}(\nabla^2 f(x)) \geq \mu$.

We also have the following lemma similar to the first order condition for convexity and smoothness.

Lemma

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Then f is μ-strongly convex iff for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x}\rangle + \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{x}\|^2 .$$

Proof

Let $g(\boldsymbol{x}) = f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|^2$. By the first order condition for convexity, $g(\boldsymbol{x})$ is convex iff for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \langle \nabla g(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle .$$

Note that $\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \mu \boldsymbol{x}$. So it gives that $g(\boldsymbol{x})$ is convex iff for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$g(\boldsymbol{y}) \geq g(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle - \mu\langle \boldsymbol{x}, \boldsymbol{y} - \boldsymbol{x}\rangle .$$

The last inequality is equivalent to

$$f(\boldsymbol{y}) - \frac{\mu}{2}\|\boldsymbol{y}\|^2 \geq f(\boldsymbol{x}) - \frac{\mu}{2}\|\boldsymbol{x}\|^2 + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle - \mu\langle \boldsymbol{x}, \boldsymbol{y}\rangle + \mu\langle \boldsymbol{x}, \boldsymbol{x}\rangle .$$

Rearranging it, we obtain that

$$\begin{aligned} f(\boldsymbol{y}) &\geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle - \mu \langle \boldsymbol{x}, \boldsymbol{y} \rangle + \frac{\mu}{2} \|\boldsymbol{x}\|^2 + \frac{\mu}{2} \|\boldsymbol{y}\|^2 \\ &= f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\mu}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 . \end{aligned}$$

As a corollary, above lemma implies that $f(y) > f(x) + \nabla f(x)^\mathsf{T}(y - x)$ for any $x \neq y$. Hence, f is strictly convex.

Example

1. An affine functions $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ can not be strongly convex since it is not strictly convex.
2. $f(x) = -\log x$ can not be strongly convex since $f'(x) = -\frac{1}{x}$ and $f''(x) = \frac{1}{x^2}$ and we can not find out such μ when $x \to 0$.
3. $f(x) = ax^2, a > 0$ is $2a$-strongly convex.
4. $f(x) = x^4$ is not strongly convex since $f'(x) = 4x^3$, $f''(x) = 12x^2$ and we can not find such $\mu > 0$.
5. $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{Q}\boldsymbol{x}$ where $\boldsymbol{Q} \succ 0$ is strongly convex. Because $\nabla^2 f(\boldsymbol{x}) = 2\boldsymbol{Q}$, f is $2\lambda_{\min}(\boldsymbol{Q})$-strongly convex.

Recall the property of monotone gradient for convex functions. We have a similar corollary.

Corollary

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. Then f is μ-strongly convex iff for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,

$$\langle \nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle \geq \mu \|\boldsymbol{x} - \boldsymbol{y}\|^2 .$$

10.4 Convergence of gradient descent with strong convexity

We now establish the convergence of gradient descent with strong convexity. First consider the gradient flow again. By strong convexity, we can bound the derivative

as follows

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\|\boldsymbol{x}_t-\boldsymbol{x}^*\|^2 &= 2\left\langle \boldsymbol{x}_t-\boldsymbol{x}^*, \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t\right\rangle \\ &= -2\left\langle \boldsymbol{x}_t-\boldsymbol{x}^*, \nabla f(\boldsymbol{x}_t)\right\rangle \\ &= -2\left\langle \boldsymbol{x}_t-\boldsymbol{x}^*, \nabla f(\boldsymbol{x}_t)-\nabla f(\boldsymbol{x}^*)\right\rangle \\ &\leq -2\mu\|\boldsymbol{x}_t-\boldsymbol{x}^*\|^2 .\end{aligned}$$

For a time-continuous non-negative process $u_t = u(t)$, if $\frac{\mathrm{d}}{\mathrm{d}t}u_t = -\alpha u_t$, then we have $u_t = u_0\exp(-\alpha t)$. The same result holds if we replace the equality by an inequality.

Theorem (*Gronwall's lemma*)

For a time-continuous non-negative process $u_t = u(t)$, if $\frac{\mathrm{d}}{\mathrm{d}t}u_t \leq -\alpha_t u_t$, then we have

$$u_T \leq u_0\exp\Big(-\int_0^T \alpha_t\,\mathrm{d}t\Big) .$$

Applying the Gronwall's lemma, we conclude $\|\boldsymbol{x}_T-\boldsymbol{x}^*\|^2 \leq \|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2\exp(-2\mu T)$ immediately, which gives an exponential decay rate. Intuitively, as the discretization version of the gradient flow, the gradient descent for strongly convex functions should also follow the exponential decay.

Theorem

Suppose $f:\mathbb{R}^n\to\mathbb{R}$ is n L-smooth and μ-strongly convex function. Choose $\eta\leq 1/L$, and let the gradient descent iterate with a fixed step size η. Then it holds that

$$\|\boldsymbol{x}_T-\boldsymbol{x}^*\|^2 \leq (1-\mu\eta)^T\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2 .$$

Proof

By strong convex,

$$\begin{aligned}\|\boldsymbol{x}_{k+1}-\boldsymbol{x}^*\|^2 &= \langle(\boldsymbol{x}_k-\boldsymbol{x}^*)-\eta\nabla f(\boldsymbol{x}_k),\,(\boldsymbol{x}_k-\boldsymbol{x}^*)-\eta\nabla f(\boldsymbol{x}_k)\rangle\\ &= \|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+\eta^2\|\nabla f(\boldsymbol{x}_k)\|^2-2\eta\langle\nabla f(\boldsymbol{x}_k),\boldsymbol{x}_k-\boldsymbol{x}^*\rangle\\ &\le \|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+\eta^2\|\nabla f(\boldsymbol{x}_k)\|^2-2\eta\Big(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)+\frac{\mu}{2}\|\boldsymbol{x}_k-\boldsymbol{x}^*\|\\ &= (1-\mu\eta)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+\eta^2\|\nabla f(\boldsymbol{x}_k)\|^2-2\eta\big(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\big)\\ &\le (1-\mu\eta)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2+2\eta\big(f(\boldsymbol{x}_k)-f(\boldsymbol{x}_{k+1})\big)-2\eta\big(f(\boldsymbol{x}_k)-f(\boldsymbol{x}^*)\\ &= (1-\mu\eta)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2-2\eta\big(f(\boldsymbol{x}_{k+1})-f(\boldsymbol{x}^*)\big)\\ &\le (1-\mu\eta)\|\boldsymbol{x}_k-\boldsymbol{x}^*\|^2\end{aligned}$$

where the second inequality is due to the *descent lemma*
$f(\boldsymbol{x}_{k+1})\le f(\boldsymbol{x}_k)-\frac{\eta}{2}\|\nabla f(\boldsymbol{x}_k)\|^2$.

The function value also has an exponential decay. Since f is L-smooth, we have

$$f(\boldsymbol{x}_T)-f(\boldsymbol{x}^*)\le\langle\nabla f(\boldsymbol{x}^*),\,\boldsymbol{x}_T-\boldsymbol{x}^*\rangle+\frac{L}{2}\|\boldsymbol{x}_T-\boldsymbol{x}^*\|^2,$$

which gives the following corollary.

Corollary

$$f(\boldsymbol{x}_T)-f^*\le\frac{L}{2}(1-\mu\eta)^T\|\boldsymbol{x}_0-\boldsymbol{x}^*\|^2.$$

10.5 Condition number

For a quadratic function $f(\boldsymbol{x})=\frac{1}{2}\boldsymbol{x}^\mathsf{T}Q\boldsymbol{x}$ where $Q\succeq 0$, we already have the following facts:

- $f(\boldsymbol{x})$ is $\lambda_{\min}(Q)$-strongly convex;
- $f(\boldsymbol{x})$ is $\lambda_{\max}(Q)$-smooth.

Applying the above theorem, if we take the step size $\eta=1/\lambda_{\max}(Q)$, $\{\boldsymbol{x}_n\}$ will converge at an exponential rate of $1-\frac{\lambda_{\min}(Q)}{\lambda_{\max}(Q)}$ (since $\boldsymbol{x}^*=\boldsymbol{0}$). Recall the in last lecture, we have shown $f(\boldsymbol{x}_{k+1})<f(\boldsymbol{x}_k)$ as long as $\eta<2/\lambda_{\max}(Q)$, which means that $\eta=1/\lambda_{\max}(Q)$ is not necessary for convergence. Since we hope $1-\mu\eta$ is as small as possible, we may choose a greater value of η to obtain a better rate.

Now let us calculate the optimal convergence rate of $f(\boldsymbol{x})$. Since $\nabla f(\boldsymbol{x}) = Q\boldsymbol{x}$, we have

$$\boldsymbol{x}_{k+1} = (I - \eta Q)\,\boldsymbol{x}_k\,.$$

Applying the eigen-decomposition $Q = U\Lambda U^\mathsf{T}$, where $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $UU^\mathsf{T} = I$.
Then $I - \eta Q = U\Lambda' U^\mathsf{T}$ where $\Lambda' = \operatorname{diag}\{1-\eta\lambda_1, 1-\eta\lambda_2, \ldots, 1-\eta\lambda_n\}$. So

$$\boldsymbol{x}_k = \left(U\Lambda' U^\mathsf{T}\right)^k \boldsymbol{x}_0 = U\left(\Lambda'\right)^k U^\mathsf{T}\boldsymbol{x}_0\,.$$

Let $\boldsymbol{y}_k = U^\mathsf{T}\boldsymbol{x}_k$. We have
$\boldsymbol{y}_k = \left(\Lambda'\right)^k \boldsymbol{y}_0 = \operatorname{diag}\left\{(1-\eta\lambda_1)^k, (1-\eta\lambda_2)^k, \ldots, (1-\eta\lambda_n)^k\right\}\boldsymbol{y}_0$.

Note that $\boldsymbol{x}^* = \arg\min f(\boldsymbol{x}) = \mathbf{0}$. Thus,

$$\begin{aligned}
\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 = \|\boldsymbol{x}_k\|^2 &= \boldsymbol{x}_k^\mathsf{T}\boldsymbol{x}_k = \boldsymbol{x}_k^\mathsf{T} U U^\mathsf{T}\boldsymbol{x}_k \\
&= \boldsymbol{y}_k^\mathsf{T}\boldsymbol{y}_k = \|\boldsymbol{y}_k\|^2 = \sum_{i=1}^n (1-\eta\lambda_i)^{2k} y_{0,i}^2 \\
&\le \max_{1\le i\le n}\left\{(1-\eta\lambda_i)^{2k}\right\}\|\boldsymbol{y}_0\|^2 \\
&= \left(\max_{1\le i\le n}\{1-\eta\lambda_i\}\right)^{2k}\|\boldsymbol{y}_0\|^2 \\
&= \max\left\{|1-\eta\lambda_{\min}|, |1-\eta\lambda_{\max}|\right\}^{2k}\|\boldsymbol{y}_0\|^2 \\
&= \max\left\{|1-\eta\lambda_{\min}|, |1-\eta\lambda_{\max}|\right\}^{2k}\|\boldsymbol{x}_0\|^2
\end{aligned}.$$

We would like to choose η to minimize $\max\{|1-\eta\lambda_{\min}|, |1-\eta\lambda_{\max}|\}$, which means $\eta = \frac{2}{\lambda_{\min}+\lambda_{\max}} = \frac{2}{\mu+L}$. In this case,

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \le \left|\frac{\lambda_{\max}-\lambda_{\min}}{\lambda_{\max}+\lambda_{\min}}\right|^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 = \left|\frac{L-\mu}{L+\mu}\right|^{2k}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2\,.$$

Definition (*Condition number*)

Given an $n \times n$ positive definite matrix $Q \succ 0$, its *condition number* is defined by

$$\kappa(Q) = \frac{\lambda_{\max}(Q)}{\lambda_{\min}(Q)} \ge 1\,.$$

The argument above reveals that for quadratic functions, the convergence rate of gradient descent depends on $\left(\frac{\kappa-1}{\kappa+1}\right)^2$.

Example

- For $Q = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$, its condition number is 2, so the convergence rate is $\frac{1}{3}$.
- For $Q = \begin{pmatrix} \frac{1}{100} & 0 \\ 0 & 1 \end{pmatrix}$, its condition number is 100, so the convergence rate is $\frac{99}{101}$.

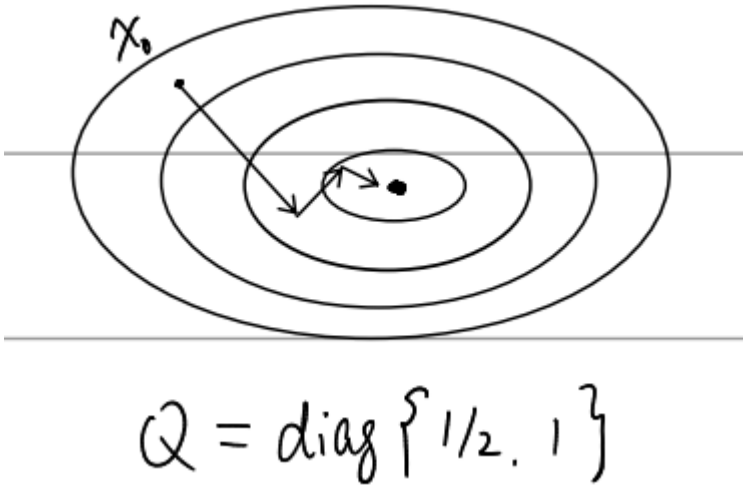
These two examples shows that the gradient descent may converge very slowly when the coefficient matrix has a large condition number.

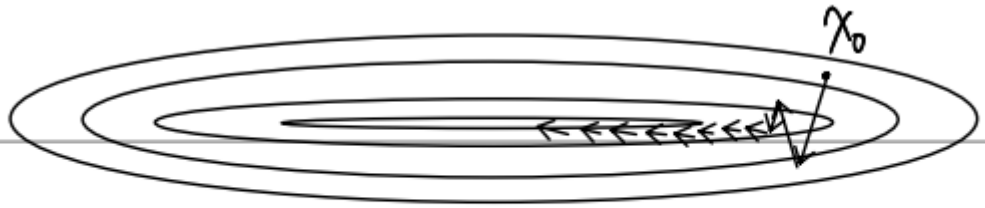
For nonquadratic functions, we can approximate them locally (near the minimum point $\boldsymbol{x}^*$) by the following Taylor series

$$\begin{aligned} f(x) &\approx f(x^*) + \nabla f(x^*)^\mathsf{T}(x - x^*) + \frac{1}{2}(x - x^*)^\mathsf{T}\nabla^2 f(x^*)(x - x^*) \\ &= f(x^*) + \frac{1}{2}(x - x^*)^\mathsf{T}\nabla^2 f(x^*)(x - x^*) \end{aligned}$$

Hence, in the neighborhood of the minimum point $\boldsymbol{x}^*$, the convergence rate depends on $\left(\frac{\kappa(\nabla^2 f(x^*))-1}{\kappa(\nabla^2 f(x^*))+1}\right)^2$. If the condition number of $\nabla^2 f(x^*)$ is large, the results given by gradient descent with fixed step size cannot converge rapidly.

Here are some well-conditioned and ill-conditioned examples:





$Q = \text{diag}\{0.01, 1\}$