# Gradient Descent and Convergence Analysis

## 1 Descent method for unconstrained optimization problems

We now study the general convex optimization problems. First, we consider the easiest case: no constraints. Namely, the optimization problem is

$$\min_{x\in\mathbb{R}^n} \quad f(x)$$

where $f(x)$ is a convex function.

Recall that, the optimality condition for convex functions is

**Theorem**

Suppose $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function. Then $x^*$ is a *global* minimum point of $f$ iff

$$\forall\, y \in D, \quad \nabla f(x^*)^\mathsf{T}(y - x^*) \geq 0\,.$$

In particular, if $D = \mathbb{R}^n$, then $x^*$ is a global minimum point iff $\nabla f(x^*) = \mathbf{0}$.

For convenience, we assume that $D = \mathbb{R}^n$, the objective function $f(x)$ is differentiable and has a finite minimum point $x^*$ (and the minimum value $f^*$). For some simple cases, we can compute the minimum point by solving the equation $\nabla f(x^*) = 0$. However, in general we cannot expect that closed-form solutions always exist. So we introduce some algorithms to find optimal solutions.

Suppose we know a solution $x$, which is not optimal. Can we find a better solution $y$? The convexity guarantees that

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T}(y - x)\,.$$

As we hope $y$ is better, i.e., $f(y) < f(x)$, it requires that $\nabla f(x)^\mathsf{T}(y - x) < 0$. Conversely, we know that if the directional derivative $\nabla f(x)^\mathsf{T} v < 0$, then there

exists $\varepsilon > 0$ such that $f(x + \varepsilon v) < f(x)$. So $\nabla f(x)^\mathsf{T} v < 0$ is a reasonable requirement for the direction $v = y - x$.

This inspired the so-called *descent method*: start from a solution $\boldsymbol{x}_0$ and move to $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + t_k \boldsymbol{v}_k$ iteratively, where $t_k$ is the *step size* to be determined and $\boldsymbol{v}_k$ is the moving direction satisfying $\nabla f(\boldsymbol{x}_k)^\mathsf{T} \boldsymbol{v}_k < 0$.

The ideal stopping criterion is $\nabla f(\boldsymbol{x}_k) = \boldsymbol{0}$ for some $k$. If so, we know that $x_k$ is indeed a minimum point. However, in practice, we cannot expect this happens. So we usually use stopping criteria such as $\|\nabla f(\boldsymbol{x})\| < \delta$, $|f(\boldsymbol{x}_{k+1}) - f(\boldsymbol{x}_k)| < \delta$, or 1000 iterations.

given a starting point $\boldsymbol{x}_0$
repeat
    choose a proper step size $t_k$
    $\boldsymbol{x}_{k+1} \leftarrow \boldsymbol{x}_k + t_k \boldsymbol{v}_k$ where $\nabla f(\boldsymbol{x}_k)^\mathsf{T} \boldsymbol{v}_k < 0$
    $k \leftarrow k + 1$
until $\|\nabla f(\boldsymbol{x}_k)\| \le \delta$ for some sufficiently small $\delta$

We now consider a specific descent method, the *gradient descent*, where we select $v_k = -\nabla f(x_k)$. Then trivially $\nabla f(x_k)^\mathsf{T} v_k < 0$.

There is an advantage to choose $-\nabla f(x_k)$ since it is the direction of *steepest descent*, namely, the value of $f$ decreases most rapidly: For any *unit* length vector $v$, the directional derivative $\nabla f(x)^\mathsf{T} v$ satisfies

$$-\|v\| \cdot \|\nabla f(x)\| \le \nabla f(x)^\mathsf{T} v \le \|v\| \cdot \|\nabla f(x)\|$$

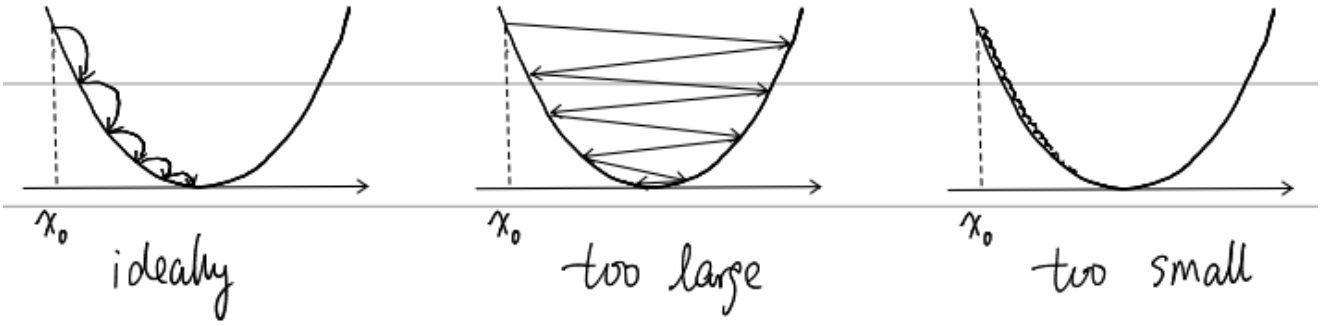
by the Cauchy-Schwarz inequality, and the equality holds iff $v = \pm \nabla f(x) / \|\nabla f(x)\|$.

Applying this choice of directions, we obtain the *gradient descent method*:

given a starting point $x_0$
repeat
    choose a proper step size $t_k$
    $x_{k+1} \leftarrow x_k - t_k \nabla f(x_k)$
    $k \leftarrow k + 1$
until $\|\nabla f(x_k)\| \le \delta$ for some sufficiently small $\delta$

The key problem is how to choose an appropriate step size $t_k$ to guarantee that the $\{x_k\}$ converges to the optimal solution. Intuitively, the choice of step size can effect

the correctness and the converge rate of the algorithm.




The easiest way is to choose a constant. Let's start from an easy example: $f(x) = ax^2$ where $a > 0$. Since we hope $f(x_{k+1}) < f(x_k)$, it requires that $|x_{k+1}| < |x_k|$, which is equivalent to

$$|(1 - 2at_k)x_k| < |x_k| \,.$$

So $t_k < 1/a$ suffices.

Next, consider the multivariate function $f(\boldsymbol{x}) = \boldsymbol{x}^\mathsf{T} Q \boldsymbol{x}$ where $Q \succeq 0$. Now $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - 2t_k Q \boldsymbol{x}_k$. So

$$f(\boldsymbol{x}_{k+1}) = \boldsymbol{x}_k^\mathsf{T} Q \boldsymbol{x}_k + 4t_k^2 (Q\boldsymbol{x}_k)^\mathsf{T} Q (Q\boldsymbol{x}_k) - 4t_k (Q\boldsymbol{x}_k)^\mathsf{T} (Q\boldsymbol{x}_k) \,.$$

It is sufficient to find a value of $t_k$ such that for all $\boldsymbol{v} \in \mathbb{R}^n$, $t_k \boldsymbol{v}^\mathsf{T} Q \boldsymbol{v} < \boldsymbol{v}^\mathsf{T} \boldsymbol{v}$. We need the following lemma.

**Lemma (*Rayleigh quotient*)**

Let $Q \succeq 0$ be a positive semi-definite matrix, and $\lambda_{\min}$ and $\lambda_{\max}$ be its minimum and maximum eigenvalues, respectively. Then for all $x \in \mathbb{R}^n$, we have

$$\lambda_{\min} \|x\|_2^2 \leq x^\mathsf{T} Q x \leq \lambda_{\max} \|x\|_2^2$$

**Proof**

Since $Q \in \mathbb{R}^{n \times n}$ is symmetric, consider its eigen-decomposition $Q = U \Lambda U^\mathsf{T}$, where $\Lambda = \mathrm{diag}\{\lambda_1, \ldots, \lambda_n\}$ is the diagonal matrix consisting of $Q$'s eigenvalues, and $U = (u_1, \ldots, u_n)$ consists of corresponding unit-length eigenvectors. It easy to see that $UU^\mathsf{T} = I$.

Assume $\boldsymbol{x} = U\boldsymbol{y}$ (i.e. $\boldsymbol{y} = U^{-1}\boldsymbol{x} = U^{\mathsf{T}}\boldsymbol{x}$). Then

$$\boldsymbol{x}^{\mathsf{T}}Q\boldsymbol{x} = \boldsymbol{y}^{\mathsf{T}}U^{\mathsf{T}}QU\boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}}U^{\mathsf{T}}U\Lambda U^{\mathsf{T}}U\boldsymbol{y} = \boldsymbol{y}^{\mathsf{T}}\Lambda\boldsymbol{y} = \sum_{i=1}^{n}\lambda_i y_i^2 \,.$$

So clearly we have $\lambda_{\min}\|\boldsymbol{y}\|^2 \leq \boldsymbol{x}^{\mathsf{T}}Q\boldsymbol{x} \leq \lambda_{\max}\|\boldsymbol{y}\|^2$. Moreover, we have

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} = \boldsymbol{x}^{\mathsf{T}}U^{\mathsf{T}}U\boldsymbol{x} = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{x} = \|\boldsymbol{x}\|^2 \,,$$

which completes the proof.

Note that in this proof we do not really need $Q \succeq 0$. This lemma holds for all symmetric $Q$. Applying this lemma, it gives that $t_k < 1/\lambda_{\max}$ suffices in the gradient descent method for quadratic functions.

However, for general cases, we cannot expect a universal condition for $t_k$. For example, consider the function $f(x) = |x|$. If we choose $t_k$ to be a constant $t > 0$, no matter what value $t$ is, the algorithm does not work as long as $|x_k| < t$.

**Question**

Under which assumptions can we choose a constant as the step size?

## 2. Smoothness

We would like to avoid functions similar to $|x|$, where $\nabla f(x)$ changes too drastically near $x^*$.

**Definition (*Lipschitz continuity*)**

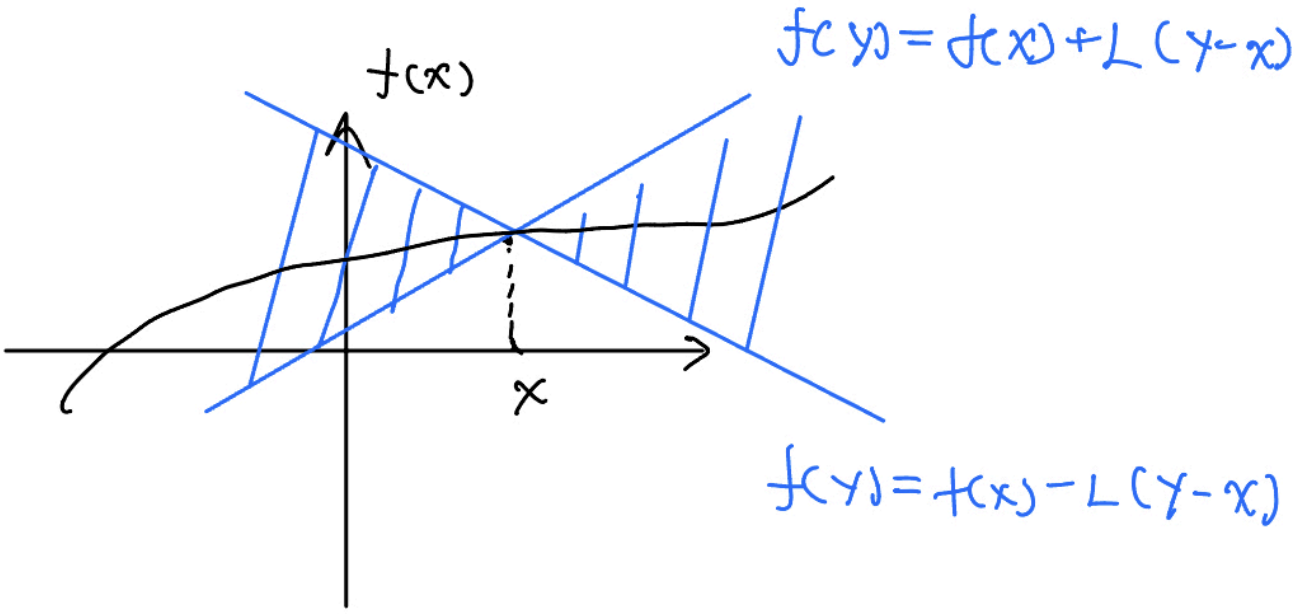
A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-Lipschitz, if for all $x, y \in \operatorname{dom} f$,

$$\|f(x) - f(y)\| \leq L\|x - y\| \,.$$

We usually use $L^2$-norm, unless otherwise specified.

An $L$-Lipschitz function is continuous, but may not be differentiable. Intuitively, for a Lipschitz continuous function, there exists a double cone (white) whose origin can be moved along the graph so that the whole graph always stays outside the

double cone.




**Example**

- $f(x) = kx$ where $x \in \mathbb{R}$ is $|k|$-Lipschitz.
- $f(\boldsymbol{x}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{x}$ where $\boldsymbol{x} \in \mathbb{R}^n$ is $\|\boldsymbol{w}\|$-Lipschitz
- $f(\boldsymbol{x}) = Q\boldsymbol{x}$ where $\boldsymbol{x} \in \mathbb{R}^n$ is $\lambda_{\max}(Q^\mathsf{T}Q)^{1/2}$-Lipschitz, since

$$\begin{aligned}\|f(\boldsymbol{x}) - f(\boldsymbol{y})\| = \|Q(\boldsymbol{x}-\boldsymbol{y})\| &= \left((\boldsymbol{x}-\boldsymbol{y})^\mathsf{T}Q^\mathsf{T}Q(\boldsymbol{x}-\boldsymbol{y})\right)^{1/2} \\ &\leq \lambda_{\max}(Q^\mathsf{T}Q)^{1/2}\|\boldsymbol{x}-\boldsymbol{y}\|\end{aligned}$$

by the bound for the Rayleigh quotient. In particular, if $Q$ is symmetric,

$$\lambda_{\max}(Q^\mathsf{T}Q)^{1/2} = \max\{|\lambda_{\min}(Q)|, |\lambda_{\max}(Q)|\}\,.$$

Recall that we hope $\nabla f(x)$ does not change rapidly. So we define the following notion of "smoothness".

**Definition (*Smoothness*)**

A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L$-smooth if $\nabla f$ if $L$-Lipschitz, i.e., for all $x, y$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|\,.$$

**Example**

$f(\boldsymbol{x}) = \boldsymbol{x}^\top Q\boldsymbol{x}$ with $Q \succeq 0$ is $2\lambda_{\max}(Q)$-smooth ($\nabla f(\boldsymbol{x}) = 2Q\boldsymbol{x}$).

We use the notation $A \succeq B$ if $A - B \succeq 0$. Then we have the following equivalent definitions.

**Lemma**

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a twice differentiable function. Then $f$ is $L$-smooth iff $-L\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}_n$ for all $\boldsymbol{x} \in \mathbb{R}^n$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Namely, for all $\boldsymbol{x} \in \mathbb{R}^n$, $\left|\lambda_i(\nabla^2 f(\boldsymbol{x}))\right| \leq L$, where $\lambda_1, \ldots, \lambda_n$ are $n$ eigenvalues.

Note that if $f : \mathbb{R} \to \mathbb{R}$, we can easily prove the " $\Longleftarrow$ " direction since the mean value theorem gives that $f'(x) - f'(y) = f''(z)(x - y)$ for some $z$. However, there is no such theorem for vector-valued functions.

**Proof**

- " $\Longleftarrow$ " direction. We would like to restrict the vector-valued function $\nabla f$ to a line. Fix any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let $\varphi : [0, 1] \to \mathbb{R}$ be a function defined by

  $$\varphi(t) = \left\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x})) \right\rangle.$$

  Then, $\varphi(1) = \langle \nabla f(\boldsymbol{y}), \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) \rangle$ and $\varphi(0) = \langle \nabla f(\boldsymbol{x}), \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}) \rangle$. By the mean value theorem, there exists $t \in [0, 1]$ such that $\varphi(1) - \varphi(0) = \varphi'(t)$. Note that

  $$\begin{aligned} \varphi'(t) &= \left\langle \nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x}),\ \nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) \right\rangle \\ &\leq \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \cdot \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\| \end{aligned}$$

  by the Cauchy-Schwarz inequality. It implies that

  $$\begin{aligned} \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\|^2 &= \varphi(1) - \varphi(0) \\ &\leq \|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \cdot \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\|, \end{aligned}$$

  which further gives that

  $$\|\nabla f(\boldsymbol{y}) - \nabla f(\boldsymbol{x})\| \leq \|\nabla^2 f(\boldsymbol{x} + t(\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x})\| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|.$$

  The last inequality follows from the third example of Lipschitz functions.

- " $\Longrightarrow$ " direction. Fix any $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^n$. Let $\psi : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a function defined by

  $$\psi(t) = \langle \nabla f(\boldsymbol{x} + t\boldsymbol{v}),\, \boldsymbol{v} \rangle \,.$$

  Then, by the Cauchy-Schwarz inequality and the $L$-smoothness, we have

  $$\begin{aligned} |\psi(t) - \psi(0)| &= \left| \langle \nabla f(\boldsymbol{x} + t\boldsymbol{v}) - \nabla f(\boldsymbol{x}),\, \boldsymbol{v} \rangle \right| \\ &\leq \|\nabla f(\boldsymbol{x} + t\boldsymbol{v}) - \nabla f(\boldsymbol{x})\| \cdot \|v\| \\ &\leq tL\|v\|^2 \,, \end{aligned}$$

  which further gives that $\left| \frac{\psi(t) - \psi(0)}{t} \right| \leq L\|\boldsymbol{v}\|^2$. Taking the limit $t \to 0$ on both sides, and applying the chain rule, we obtain that

  $$\left| \boldsymbol{v}^\mathsf{T} \nabla^2 f(\boldsymbol{x}) \boldsymbol{v} \right| = |\psi'(0)| \leq L\|\boldsymbol{v}\|^2 \,.$$

  Thus, $-L\boldsymbol{I}_n \preceq \nabla^2 f(\boldsymbol{x}) \preceq L\boldsymbol{I}_n$.

An $L$-smooth functions may be not convex. If $f$ is further convex, all absolute values are not necessary.

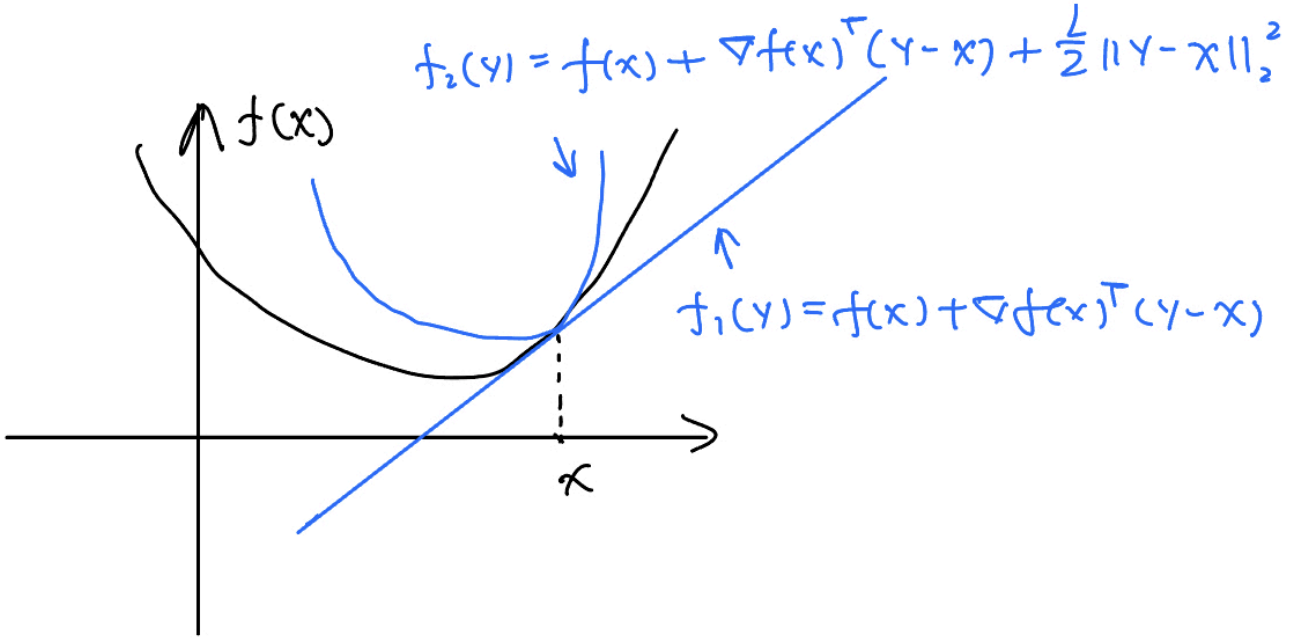
**Lemma**

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

$$\left| f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle \right| \leq \frac{L}{2} \|\boldsymbol{y} - \boldsymbol{x}\|^2 \,.$$

Recall that $f$ is convex iff $f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y} - \boldsymbol{x} \rangle \geq 0$, which shows that $f$ is underestimated by an affine function. Now, if $f$ is $L$-smooth, it is overestimated by

a quadratic function.




**Proof**

- " $\Longleftarrow$ " direction. Fix $\boldsymbol{x} \in \mathbb{R}^n$. Define

$$g_1(\boldsymbol{y}) = f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y}-\boldsymbol{x}\rangle + \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2,$$
$$g_2(\boldsymbol{y}) = f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y}-\boldsymbol{x}\rangle - \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2.$$

  Note that for all $\boldsymbol{y} \in \mathbb{R}^n$, $g_2(\boldsymbol{y}) \le 0 \le g_1(\boldsymbol{y})$, and $g_1(\boldsymbol{x}) = g_2(\boldsymbol{x}) = 0$. So $\boldsymbol{x}$ is a local minimum point of $g_1$, which gives that $\nabla^2 g_1(\boldsymbol{x}) \succeq 0$. Since $\nabla^2 g_1(\boldsymbol{y}) = \nabla^2 f(\boldsymbol{y}) + L\boldsymbol{I}_n$, we conclude that $\nabla^2 f(\boldsymbol{x}) \succeq -L\boldsymbol{I}_n$. Similarly, $\boldsymbol{x}$ is a local maximum point of $g_2$, and thus $\nabla^2 g_2(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - L\boldsymbol{I}_n \preceq 0$.
- " $\Longrightarrow$ " direction. Fix $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Let

$$h(\theta) = f(\boldsymbol{x} + \theta(\boldsymbol{y}-\boldsymbol{x})).$$

  It is clear that $h'(\theta) = \langle \nabla f(\boldsymbol{x} + \theta(\boldsymbol{y}-\boldsymbol{x})), \boldsymbol{y}-\boldsymbol{x}\rangle$, and

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) = h(1) - h(0) = \int_0^1 h'(\theta)\,\mathrm{d}\theta.$$

  Moreover, $\langle \nabla f(\boldsymbol{x}), \boldsymbol{y}-\boldsymbol{x}\rangle = h'(0) = \int_0^1 h'(0)\,\mathrm{d}\theta$. Therefore, it holds that

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) - \langle \nabla f(\boldsymbol{x}),\, \boldsymbol{y}-\boldsymbol{x}\rangle = \int_0^1 h'(\theta) - h'(0)\,\mathrm{d}\theta.$$

Note that

$$\begin{aligned}\left|h'(\theta)-h'(0)\right| &= \left|\langle \nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x}),\, \boldsymbol{y}-\boldsymbol{x}\rangle\right| \\ &\le \|\nabla f(\boldsymbol{x}+\theta(\boldsymbol{y}-\boldsymbol{x}))-\nabla f(\boldsymbol{x})\|\cdot\|\boldsymbol{y}-\boldsymbol{x}\| \\ &\le \theta L\|\boldsymbol{y}-\boldsymbol{x}\|^2 .\end{aligned}$$

We now have

$$\begin{aligned}\left|\langle f(\boldsymbol{y})-f(\boldsymbol{x})-\nabla f(\boldsymbol{x}),\, \boldsymbol{y}-\boldsymbol{x}\rangle\right| &\le \int_0^1 \left|h'(\theta)-h'(0)\right| \mathrm{d}\theta \\ &\le \int_0^1 \theta L\|\boldsymbol{y}-\boldsymbol{x}\|^2 \,\mathrm{d}\theta = \frac{L}{2}\|\boldsymbol{y}-\boldsymbol{x}\|^2 ,\end{aligned}$$

which completes the proof.

Recall that, we hope to find the value of the step size $t$ such that $f(\boldsymbol{x}_{k+1}) < f(\boldsymbol{x}_k)$. Now we assume that $f$ is $L$-smooth. Then

$$\begin{aligned}f(\boldsymbol{x}_{k+1}) &= f(\boldsymbol{x}_k - t\cdot\nabla f(\boldsymbol{x}_k)) \\ &\le f(\boldsymbol{x}_k) - \langle \nabla f(\boldsymbol{x}_k),\, t\cdot\nabla f(\boldsymbol{x}_k)\rangle + \frac{L}{2}\|t\cdot\nabla f(\boldsymbol{x}_k)\|^2 \\ &= f(\boldsymbol{x}_k) - t\Big(1-\frac{Lt}{2}\Big)\|\nabla f(\boldsymbol{x}_k)\|^2 \\ &< f(\boldsymbol{x}_k)\end{aligned}$$

if we set $t < 2/L$. In particular, if we choose $t \le 1/L$, it gives the following *descent lemma*.

**Lemma (*Descent lemma*)**

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

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

## 3. Convergence analysis with smoothness

We first introduce a continuous version of the gradient descent instead, which is easier to analyze.

**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

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

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

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

We compare the gradient descent with the gradient flow. Assume the gradient descent iterates with a fixed step size $\eta$ 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 $\nabla f$ 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 $\eta \leq 1/L$, and let the gradient descent iterate with a fixed step size $\eta$. Then it holds that

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

**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

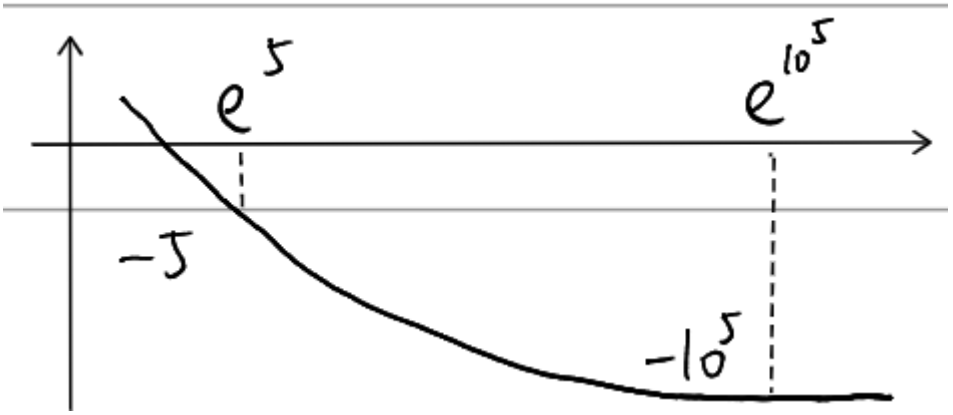
$$\|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \leq \sum_{k=1}^{T} 2\eta\left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_k)\right) \leq 2T\eta\left(f(\boldsymbol{x}^*) - f(\boldsymbol{x}_T)\right),$$

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 $\varepsilon$ 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?

## 4. Strong convexity

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 $x^\top Q x$ 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, $x^\top Q x$ is almost equivalent to $\lambda_{\max}(Q) x^\top x = \lambda_{\max}(Q)\|x\|^2$. 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

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

**Lemma**

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

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

## 5. Convergence analysis 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

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

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 $\mu$-strongly convex function. Choose $\eta\leq 1/L$, and let the gradient descent iterate with a fixed step size $\eta$. Then it holds that

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

**Proof**

By strong convexity,

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

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**

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