# Gradient Descent

## Optimization methods in machine learning

Aleksandr Beznosikov

Innopolis University

5 September 2023

# Algorithm

- **Goal:** Find $x \in \mathbb{R}^d$ such that

$$f(x) - f(x^*) \leq \varepsilon.$$

- Assumptions: $f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, an optimal value $f^*$ exists (remember $\min_{x \in \mathbb{R}} x$) and a unique global minima $x^*$ also exists
- Note that there can be several minima $x_1^* \neq x_2^*$ with $f(x_1^*) = f(x_2^*)$.
  **Question:** Can you give me such an non-trivial example of function?

# Algorithm

- **Goal:** Find $\mathrm{x} \in \mathbb{R}^d$ such that

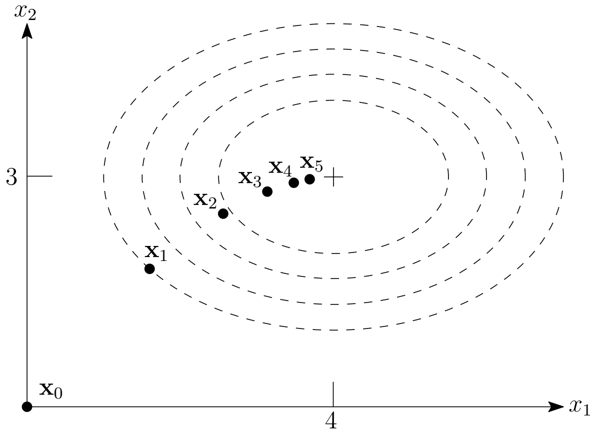
$$f(\mathrm{x}) - f(\mathrm{x}^*) \leq \varepsilon.$$

- Assumptions: $f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, an optimal value $f^*$ exists (remember $\min_{x \in \mathbb{R}} x$) and a unique global minima $x^*$ also exists
- Note that there can be several minima $\mathrm{x}_1^* \neq \mathrm{x}_2^*$ with $f(\mathrm{x}_1^*) = f(\mathrm{x}_2^*)$.
  **Question:** Can you give me such an non-trivial example of function?
  $f(x_1, x_2) = (x_1 - x_2)^2$, all points $x_1 = x_2$ give $f = 0$.
- **Iterative Algorithm:**

$$\mathrm{x}_{k+1} := \mathrm{x}_k - \gamma \nabla f(\mathrm{x}_k),$$

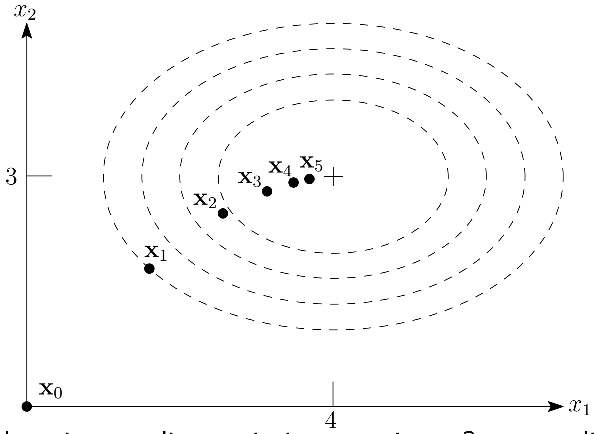
for timesteps $k = 0, 1, \ldots,$ and stepsize $\gamma \geq 0$.

# Example




**Question:** where is a gradient pointing at point $x_1$?

# Example




**Question:** where is a gradient pointing at point $x_1$? ascent direction

# Vanilla analysis

How to bound $f(x_{output}) - f(x^*)$ ?

# Vanilla analysis

How to bound $f(x_{output}) - f(x^*)$ ?

- Using the definition/update of gradient descent:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \gamma \nabla f(x_k) - x^*\|^2 = \|x_k - x^* - \gamma \nabla f(x_k)\|^2.$$

# Vanilla analysis

How to bound $f(x_{output}) - f(x^*)$ ?

- Using the definition/update of gradient descent:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \gamma \nabla f(x_k) - x^*\|^2 = \|x_k - x^* - \gamma \nabla f(x_k)\|^2.$$

- Apply $\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\langle v, w \rangle$ to rewrite

# Vanilla analysis

How to bound $f(x_{output}) - f(x^*)$ ?

- Using the definition/update of gradient descent:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \gamma \nabla f(x_k) - x^*\|^2 = \|x_k - x^* - \gamma \nabla f(x_k)\|^2.$$

- Apply $\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\langle v, w \rangle$ to rewrite

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma \langle \nabla f(x_k), x_k - x^* \rangle + \gamma^2 \|\nabla f(x_k)\|^2.$$

## Vanilla analysis

How to bound $f(x_{output}) - f(x^*)$ ?

- Using the definition/update of gradient descent:

$$\|x_{k+1} - x^*\|^2 = \|x_k - \gamma \nabla f(x_k) - x^*\|^2 = \|x_k - x^* - \gamma \nabla f(x_k)\|^2.$$

- Apply $\|v - w\|^2 = \|v\|^2 + \|w\|^2 - 2\langle v, w\rangle$ to rewrite

$$\|x_{k+1} - x^*\|^2 = \|x_k - x^*\|^2 - 2\gamma\langle \nabla f(x_k), x_k - x^*\rangle + \gamma^2\|\nabla f(x_k)\|^2.$$

- Rearrangement:

$$\langle \nabla f(x_k), x_k - x^*\rangle = \frac{1}{2\gamma}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) + \frac{\gamma}{2}\|\nabla f(x_k)\|^2.$$

# Vanilla analysis

- Sum this up over the iterations $k$:

$$\sum_{k=0}^{K-1} \langle \nabla f(x_k), x_k - x^* \rangle$$
$$= \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \left( \|x_0 - x^*\|^2 - \|x_K - x^*\|^2 \right)$$

- Now we invoke convexity of $f$ with $x = x_k, y = x^*$:

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle$$

giving

$$\sum_{k=0}^{K-1} \left( f(x_k) - f(x^*) \right) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2,$$

an upper bound for the <u>average error</u> $f(x_k) - f(x^*)$ over the steps

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Assume that all gradients of $f$ are bounded in norm ($\|\nabla f(\mathrm{x})\| \leq B$).

- Equivalent to $f$ being Lipschitz ($|f(x) - f(y)| \leq B\|x - y\|^2$)

**Theorem**

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathrm{x}^*$; furthermore, suppose that $\|\mathrm{x}_0 - \mathrm{x}^*\| \leq R$ and $\|\nabla f(\mathrm{x})\| \leq B$ for all $\mathrm{x}$. Choosing the stepsize $\gamma := \frac{R}{B\sqrt{K}}$, gradient descent yields $\frac{1}{K}\sum_{k=0}^{K-1} f(\mathrm{x}_k) - f(\mathrm{x}^*) \leq \frac{RB}{\sqrt{K}}$.*

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2 \leq$$

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2 \leq \frac{\gamma}{2} B^2 K + \frac{1}{2\gamma} R^2$$

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\|x_0 - x^*\|^2 \leq \frac{\gamma}{2} B^2 K + \frac{1}{2\gamma} R^2$$

- **Question:** How to choose $\gamma$?

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2 \leq \frac{\gamma}{2} B^2 K + \frac{1}{2\gamma} R^2$$

- **Question:** How to choose $\gamma$? Choose $\gamma$ such that

$$q(\gamma) = \frac{\gamma}{2} B^2 K + \frac{R^2}{2\gamma}.$$

is minimized. **Question:** Any help how to do it?

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2 \leq \frac{\gamma}{2} B^2 K + \frac{1}{2\gamma} R^2$$

- **Question:** How to choose $\gamma$? Choose $\gamma$ such that

$$q(\gamma) = \frac{\gamma}{2} B^2 K + \frac{R^2}{2\gamma}.$$

is minimized. **Question:** Any help how to do it? Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{K}}$, and $q(R/(B\sqrt{K})) = RB\sqrt{K}$.

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

## Proof.

- Plug $\|x_0 - x^*\| \leq R$ and $\|\nabla f(x_k)\| \leq B$ into Vanilla Analysis:

$$\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma} \|x_0 - x^*\|^2 \leq \frac{\gamma}{2} B^2 K + \frac{1}{2\gamma} R^2$$

- **Question:** How to choose $\gamma$? Choose $\gamma$ such that

$$q(\gamma) = \frac{\gamma}{2} B^2 K + \frac{R^2}{2\gamma}.$$

is minimized. **Question:** Any help how to do it? Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{K}}$, and $q(R/(B\sqrt{K})) = RB\sqrt{K}$. Dividing by $K$, the result follows.

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

- Recall:
$$\frac{1}{K}\sum_{k=0}^{K-1} f(\mathrm{x}_k) - f(\mathrm{x}^*) \leq \frac{RB}{\sqrt{K}}.$$

**Question:** How can we actually find a point $\tilde{x}$ such that $f(\tilde{x}) - f(x^*) \leq \frac{RB}{\sqrt{K}}$?

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

- Recall:

$$\frac{1}{K}\sum_{k=0}^{K-1} f(x_k) - f(x^*) \leq \frac{RB}{\sqrt{K}}.$$

**Question:** How can we actually find a point $\tilde{x}$ such that $f(\tilde{x}) - f(x^*) \leq \frac{RB}{\sqrt{K}}$? Jensen's inequality: $f\left(\frac{1}{K}\sum_{k=0}^{K-1} x_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(x_k)$. And we get

$$\text{average error } \leq \frac{RB}{\sqrt{K}} \leq \varepsilon$$

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

- Recall:

$$\frac{1}{K}\sum_{k=0}^{K-1} f(x_k) - f(x^*) \leq \frac{RB}{\sqrt{K}}.$$

**Question:** How can we actually find a point $\tilde{x}$ such that $f(\tilde{x}) - f(x^*) \leq \frac{RB}{\sqrt{K}}$? Jensen's inequality: $f\left(\frac{1}{K}\sum_{k=0}^{K-1} x_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(x_k)$. And we get

$$\text{average error } \leq \frac{RB}{\sqrt{K}} \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{R^2B^2}{\varepsilon^2}$$

# Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

- Recall:

$$\frac{1}{K}\sum_{k=0}^{K-1} f(\mathrm{x}_k) - f(\mathrm{x}^*) \leq \frac{RB}{\sqrt{K}}.$$

**Question:** How can we actually find a point $\tilde{x}$ such that $f(\tilde{x}) - f(x^*) \leq \frac{RB}{\sqrt{K}}$? Jensen's inequality: $f\left(\frac{1}{K}\sum_{k=0}^{K-1} \mathrm{x}_k\right) \leq \frac{1}{K}\sum_{k=0}^{K-1} f(\mathrm{x}_k)$. And we get

$$\text{average error } \leq \frac{RB}{\sqrt{K}} \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{R^2B^2}{\varepsilon^2}$$

**Question:** What is the pros and cons of this result?

# Smooth functions

## Definition

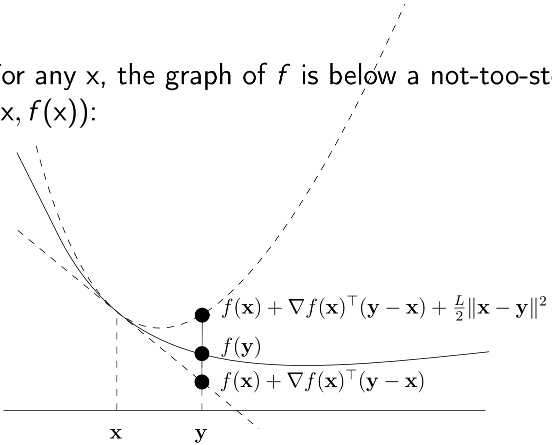
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. $f$ is called smooth (with parameter $L \geq 0$) if

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

- Definition does not require convexity (useful later)
- Definition tell us that $f$ is "not too curved".

# Smooth functions

Smoothness: For any x, the graph of $f$ is below a not-too-steep tangential paraboloid at $(x, f(x))$:



# Smooth functions: $\mathcal{O}(1/\varepsilon)$ steps

- Quadratic functions are smooth
- Operations that preserve smoothness:

## Lemma

(i) *Let $f_1, f_2, \dots, f_m$ be convex functions that are smooth with parameters $L_1, L_2, \dots, L_m$, and let $\lambda_1, \lambda_2, \dots, \lambda_m \in \mathbb{R}_+$. Then the convex function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.*

(ii) *Let $f$ be convex and smooth with parameter $L$, and let $g(\mathrm{x}) = A\mathrm{x} + \mathrm{b}$, for $A \in \mathbb{R}^{d\times m}$ and $\mathrm{b} \in \mathbb{R}^d$. Then the convex function $f \circ g$ is smooth with parameter $L\|A\|^2$, where $\|A\| = \max_{\mathrm{x}\neq 0} \frac{\|A\mathrm{x}\|}{\|\mathrm{x}\|}$ is the 2-norm (or spectral norm) of $A$.*

# Proof of lemmas

Let $f_1, f_2, \ldots, f_m$ be convex functions that are smooth with parameters $L_1, L_2, \ldots, L_m$, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the convex function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.

# Proof of lemmas

Let $f$ be convex and smooth with parameter $L$, and let $g(x) = Ax + b$, for $A \in \mathbb{R}^{d\times m}$ and $b \in \mathbb{R}^d$. Then the convex function $f \circ g$ is smooth with parameter $L\|A\|^2$, where

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

is the 2-norm (or spectral norm) of $A$.

# Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of $f$
- Smoothness $\Leftrightarrow$ Lipschitz continuity of $\nabla f$ (in the convex case).

# Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of $f$
- Smoothness $\Leftrightarrow$ Lipschitz continuity of $\nabla f$ (in the convex case).

**Lemma**

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.*

(i) *$f$ is smooth with parameter $L$.*

(ii) $\|\nabla f(\mathrm{x}) - \nabla f(\mathrm{y})\| \leq L\|\mathrm{x} - \mathrm{y}\|$ *for all* $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$.

# Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of $f$
- Smoothness $\Leftrightarrow$ Lipschitz continuity of $\nabla f$ (in the convex case).

## Lemma

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.*

(i) *$f$ is smooth with parameter $L$.*

(ii) $\|\nabla f(\mathrm{x}) - \nabla f(\mathrm{y})\| \leq L\|\mathrm{x} - \mathrm{y}\|$ *for all* $\mathrm{x}, \mathrm{y} \in \mathbb{R}^d$.

Proof in Nesterov's book (see Lemma 1.2.3).

# Proof of decreasing

- Just use smoothness for points $x_{k+1}$ and $x_k$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

- Substituting of $x_{k+1}$ from gradient descent update:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma\langle \nabla f(x_k), \nabla f(x_k)\rangle + \frac{L\gamma^2}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\gamma - \frac{\gamma^2 L}{2}\right)\|\nabla f(x_k)\|^2. \end{aligned}$$

# Proof of decreasing

- Just use smoothness for points $x_{k+1}$ and $x_k$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

- Substituting of $x_{k+1}$ from gradient descent update:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma \langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{L\gamma^2}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\gamma - \frac{\gamma^2 L}{2}\right)\|\nabla f(x_k)\|^2. \end{aligned}$$

- **Question:** what we want next?

# Proof of decreasing

- Just use smoothness for points $x_{k+1}$ and $x_k$:

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2.$$

- Substituting of $x_{k+1}$ from gradient descent update:

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) - \gamma\langle \nabla f(x_k), \nabla f(x_k) \rangle + \frac{L\gamma^2}{2}\|\nabla f(x_k)\|^2 \\ &= f(x_k) - \left(\gamma - \frac{\gamma^2 L}{2}\right)\|\nabla f(x_k)\|^2. \end{aligned}$$

- **Question:** what we want next? $\left(\gamma - \frac{\gamma^2 L}{2}\right) > 0$ – to get $f(x_{k+1}) < f(x_k)$.

# Proof of decreasing

- For the previous slide

$$f(x_{k+1}) \leq f(x_k) - \left(\gamma - \frac{\gamma^2 L}{2}\right) \|\nabla f(x_k)\|^2.$$

- **Question:** what $\gamma$ we need to $f(x_{k+1}) < f(x_k)$? what choice is optimal?

# Proof of decreasing

- For the previous slide

$$f(x_{k+1}) \leq f(x_k) - \left(\gamma - \frac{\gamma^2 L}{2}\right) \|\nabla f(x_k)\|^2.$$

- **Question:** what $\gamma$ we need to $f(x_{k+1}) < f(x_k)$? what choice is optimal? $\gamma \in (0; 2/L)$, $\gamma_{opt} = 1/L$ (just to optimize $\left(\gamma - \frac{\gamma^2 L}{2}\right)$).
- We also get

$$(2 - \gamma L)\, \frac{\gamma}{2} \|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1}).$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

## Theorem

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $x^*$; furthermore, suppose that $f$ is smooth with parameter $L$. Choosing stepsize*

$$\gamma < \frac{2}{L},$$

*gradient descent yields*

$$\sum_{k=0}^{K-1} \left(f(x_k) - f(x^*)\right) \leq \frac{1}{(2-\gamma L)} \left(f(x_0) - f(x_K)\right) + \frac{1}{2\gamma} \|x_0 - x^*\|^2$$
$$\leq \frac{1}{(2-\gamma L)} \left(f(x_0) - f(x^*)\right) + \frac{1}{2\gamma} \|x_0 - x^*\|^2, \quad K > 0.$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Proof.

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

## Proof.

Vanilla Analysis:

$$\sum_{k=0}^{K-1} \left(f(x_k) - f(x^*)\right) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\|x_0 - x^*\|^2.$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

## Proof.

Vanilla Analysis:

$$\sum_{k=0}^{K-1} \left(f(x_k) - f(x^*)\right) \leq \frac{\gamma}{2} \sum_{k=0}^{K-1} \|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\|x_0 - x^*\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\begin{aligned}\sum_{k=0}^{K-1} \left(f(x_k) - f(x^*)\right) &\leq \frac{1}{(2-\gamma L)} \sum_{k=0}^{K-1} \left(f(x_k) - f(x_{k+1})\right) + \frac{1}{2\gamma}\|x_0 - x^*\|^2 \\ &= \frac{1}{(2-\gamma L)} \left(f(x_0) - f(x_K)\right) + \frac{1}{2\gamma}\|x_0 - x^*\|^2.\end{aligned}$$

□

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Putting it together with $\gamma = 1/L$:

$$
\begin{aligned}
\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) &\leq f(x_0) - f(x_K) + \frac{L}{2}\|x_0 - x^*\|^2 \\
&\leq f(x_0) - f(x^*) + \frac{L}{2}\|x_0 - x^*\|^2.
\end{aligned}
$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}\sum_{k=0}^{K-1}(f(x_k) - f(x^*)) &\leq f(x_0) - f(x_K) + \frac{L}{2}\|x_0 - x^*\|^2 \\ &\leq f(x_0) - f(x^*) + \frac{L}{2}\|x_0 - x^*\|^2.\end{aligned}$$

Rewriting:

$$\sum_{k=1}^{K-1}(f(x_k) - f(x^*)) \leq \frac{L}{2}\|x_0 - x^*\|^2.$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}\sum_{k=0}^{K-1} (f(x_k) - f(x^*)) &\leq f(x_0) - f(x_K) + \frac{L}{2}\|x_0 - x^*\|^2 \\ &\leq f(x_0) - f(x^*) + \frac{L}{2}\|x_0 - x^*\|^2.\end{aligned}$$

Rewriting:

$$\sum_{k=1}^{K-1} (f(x_k) - f(x^*)) \leq \frac{L}{2}\|x_0 - x^*\|^2.$$

As last iterate is the best (sufficient decrease!):

$$f(x_K) - f(x^*) \leq \frac{1}{K-1}\left(\sum_{k=1}^{K-1} (f(x_k) - f(x^*))\right) \leq \frac{L}{2(K-1)}\|x_0 - x^*\|^2.$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

- Recall:

$$f(x_K) - f(x^*) \leq \frac{L}{2(K-1)} \|x_0 - x^*\|^2.$$

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error} \leq \frac{L}{2(K-1)} R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq 1 + \frac{R^2 L}{2\varepsilon}.$$

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

- Recall:

$$f(x_K) - f(x^*) \leq \frac{L}{2(K-1)} \|x_0 - x^*\|^2.$$

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error} \leq \frac{L}{2(K-1)} R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq 1 + \frac{R^2 L}{2\varepsilon}.$$

- $50 \cdot R^2 L$ iterations for error $0.01$ ...

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

- Recall:

$$f(x_K) - f(x^*) \leq \frac{L}{2(K-1)} \|x_0 - x^*\|^2.$$

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error } \leq \frac{L}{2(K-1)} R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq 1 + \frac{R^2 L}{2\varepsilon}.$$

- $50 \cdot R^2 L$ iterations for error $0.01$ ...
- ... as opposed to $10,000 \cdot R^2 B^2$ in the Lipschitz case

# Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

- Recall:

$$f(x_K) - f(x^*) \leq \frac{L}{2(K-1)} \|x_0 - x^*\|^2.$$

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error } \leq \frac{L}{2(K-1)} R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq 1 + \frac{R^2 L}{2\varepsilon}.$$

- $50 \cdot R^2 L$ iterations for error $0.01$ ...
- ...as opposed to $10,000 \cdot R^2 B^2$ in the Lipschitz case

In Practice:

What if we don't know the smoothness parameter $L$?

# Can we go even faster?

So far: Error decreases with $1/\sqrt{K}$, or $1/K$...

Could it decrease exponentially in $K$?

# Can we go even faster?

**Question:** What is $L$? What should be $\gamma$?

- On $f(x) := x^2$: Stepsize $\gamma :=$

# Can we go even faster?

**Question:** What is $L$? What should be $\gamma$?

- On $f(x) := x^2$: Stepsize $\gamma := \frac{1}{2}$ ($f$ is $L = 2$ - smooth)

$$x_{k+1} = x_k - \frac{1}{2}\nabla f(x_k) = x_k - x_k = 0,$$

  - converged in one step!

# Can we go even faster?

**Question:** What is $L$? What should be $\gamma$?

- On $f(x) := x^2$: Stepsize $\gamma := \frac{1}{2}$ ($f$ is $L = 2$ - smooth)

$$x_{k+1} = x_k - \frac{1}{2}\nabla f(x_k) = x_k - x_k = 0,$$

  - converged in one step!
- Same $f(x) := x^2$: Stepsize $\gamma := \frac{1}{4}$ ($f$ is $L = 4$ - smooth) **Question:** am I correct here?

$$x_{k+1} =$$

# Can we go even faster?

**Question:** What is $L$? What should be $\gamma$?

- On $f(x) := x^2$: Stepsize $\gamma := \frac{1}{2}$ <u>($f$ is $L = 2$ - smooth)</u>

$$x_{k+1} = x_k - \frac{1}{2}\nabla f(x_k) = x_k - x_k = 0,$$

  - converged in one step!
- Same $f(x) := x^2$: Stepsize $\gamma := \frac{1}{4}$ <u>($f$ is $L = 4$ - smooth)</u> **Question:** am I correct here?

$$x_{k+1} = x_k - \frac{1}{4}\nabla f(x_k) = x_k - \frac{x_k}{2} = \frac{x_k}{2},$$

so $f(x_k) = f\left(\frac{x_0}{2^k}\right) = \frac{1}{2^{2k}}x_0^2$.
  - Exponential in $k$ !

# Strongly convex functions

## Definition

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathrm{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function $f$ is called strongly convex (with parameter $\mu$) over $X$ if

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

# Strongly convex functions

## Definition

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathrm{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function $f$ is called strongly convex (with parameter $\mu$) over $X$ if

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

# Strongly convex functions

## Definition

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathrm{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function $f$ is called strongly convex (with parameter $\mu$) over $X$ if

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

It means that the function is "Not too flat".

# Strongly convex functions

## Definition

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathrm{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function $f$ is called strongly convex (with parameter $\mu$) over $X$ if

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

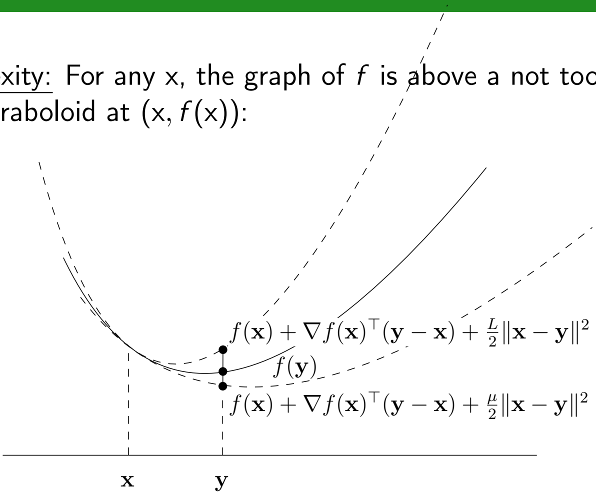
It means that the function is "Not too flat".

## Lemma

*If $f$ is strongly convex with parameter $\mu > 0$, then $f$ is strictly convex and has a unique global minimum.*

# Strongly convex functions II

Strong convexity: For any x, the graph of $f$ is above a not too flat tangential paraboloid at $(x, f(x))$:



# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Since solution is unique, we want to show: $\lim_{k\to\infty} x_k = x^*$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Since solution is unique, we want to show: $\lim_{k\to\infty} x_k = x^*$
Vanilla Analysis:

$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Since solution is unique, we want to show: $\lim_{k\to\infty} x_k = x^*$
Vanilla Analysis:

$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

Now use stronger lower bound on left hand side, coming from strong convexity:

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

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Since solution is unique, we want to show: $\lim_{k\to\infty} x_k = x^*$

Vanilla Analysis:

$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

Now use stronger lower bound on left hand side, coming from strong convexity:

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

Putting it together:

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x_k - x^*\|^2$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Since solution is unique, we want to show: $\lim_{k\to\infty} x_k = x^*$

Vanilla Analysis:

$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{\gamma}{2}\|\nabla f(x_k)\|^2 + \frac{1}{2\gamma}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

Now use <u>stronger</u> lower bound on left hand side, coming from <u>strong</u> convexity:

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

Putting it together:

$$f(x_k) - f(x^*) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) - \frac{\mu}{2}\|x_k - x^*\|^2$$

Rewriting:

$$\|x_{k+1} - x^*\|^2 \leq 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 + (1 - \mu\gamma)\|x_k - x^*\|^2.$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

$$\underline{\|x_{k+1} - x^*\|^2} \leq 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 + \underline{(1-\mu\gamma)\|x_k - x^*\|^2}.$$

Squared distance to $x^*$ goes down by a constant factor, up to some "noise".

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

$$\underline{\|x_{k+1} - x^*\|^2} \leq 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 + \underline{(1 - \mu\gamma)\|x_k - x^*\|^2}.$$

Squared distance to $x^*$ goes down by a constant factor, up to some "noise".

**Theorem**

*Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with a global minimum $x^*$; suppose that $f$ is smooth with parameter $L$ and strongly convex with parameter $\mu > 0$. Choosing $\gamma := \frac{1}{L}$, gradient descent with arbitrary $x_0$ satisfies the following property:*

- *Squared distances to $x^*$ are geometrically decreasing:*

$$\|x_{k+1} - x^*\|^2 \leq (1 - \mu\gamma)\|x_k - x^*\|^2, \quad k \geq 0.$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

$$\underline{\|x_{k+1} - x^*\|^2} \leq 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 + \underline{(1 - \mu\gamma)\|x_k - x^*\|^2}.$$

**Proof.**

Bounding the noise:

$$\begin{aligned} 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 &= 2\gamma(f(x_{k+1}) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 \\ &\leq -\gamma^2((2 - \gamma L)\|\nabla f(x_k)\|^2) + \gamma^2\|\nabla f(x_k) \end{aligned}$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

$$\underline{\|x_{k+1} - x^*\|^2} \leq 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 + \underline{(1-\mu\gamma)\|x_k - x^*\|^2}.$$

## Proof.

Bounding the noise:

$$\begin{aligned} 2\gamma(f(x^*) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 &= 2\gamma(f(x_{k+1}) - f(x_k)) + \gamma^2\|\nabla f(x_k)\|^2 \\ &\leq -\gamma^2((2-\gamma L)\,\|\nabla f(x_k)\|^2) + \gamma^2\|\nabla f(x_k) \end{aligned}$$

Hence, with $\gamma \leq \frac{1}{L}$ the noise is nonpositive, and we get:

$$\|x_{k+1} - x^*\|^2 \leq (1-\mu\gamma)\|x_k - x^*\|^2.$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error} \;\leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^K R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right).$$

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error} \;\leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^K R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right).$$

- **Conclusion:** To reach absolute error at most $\varepsilon$, we only need $\mathcal{O}(\log \frac{1}{\varepsilon})$ iterations, e.g.
  - $\frac{L}{\mu} \ln(50 \cdot R^2 L)$ iterations for error $0.01$ ...

# Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

- $R^2 := \|x_0 - x^*\|^2$.

$$\text{error} \ \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^K R^2 \leq \varepsilon \quad \Rightarrow \quad K \geq \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right).$$

- **Conclusion:** To reach absolute error at most $\varepsilon$, we only need $\mathcal{O}(\log \frac{1}{\varepsilon})$ iterations, e.g.
  - $\frac{L}{\mu} \ln(50 \cdot R^2 L)$ iterations for error 0.01 ...
  - ... as opposed to $50 \cdot R^2 L$ in the smooth case