

# Optimization for Machine Learning

## CS-439

Lecture 5: Stochastic Gradient Descent and Non-convex optimization

**Nicolas Flammarion**

EPFL – [github.com/epfml/0ptML\\_course](https://github.com/epfml/0ptML_course)

March 20, 2020

# Chapter 5

## Stochastic Gradient Descent

# Stochastic gradient descent

Many objective functions are **sum structured**:

$$f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}).$$

Example:  $f_i$  is the cost function of the  $i$ -th observation, taken from a training set of  $n$  observation.

Evaluating  $\nabla f(\mathbf{x})$  of a sum-structured function is expensive (sum of  $n$  gradients).

# Stochastic gradient descent: the algorithm

choose  $\mathbf{x}_0 \in \mathbb{R}^d$ .

sample  $i \in [n]$  uniformly at random

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \nabla f_i(\mathbf{x}_t).$$

for **times**  $t = 0, 1, \dots$ , and **stepsizes**  $\gamma_t \geq 0$ .

Only update with the gradient of  $f_i$  instead of the full gradient!

Iteration is  $n$  times cheaper than in full gradient descent.

The vector  $\mathbf{g}_t := \nabla f_i(\mathbf{x}_t)$  is called a **stochastic gradient**.

$\mathbf{g}_t$  is a vector of  $d$  random variables, but we will also simply call this a random variable.

# Unbiasedness

Can't use convexity

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)$$

on top of the vanilla analysis, as this may hold or not hold, depending on how the stochastic gradient  $\mathbf{g}_t$  turns out.

We will show (and exploit): the inequality holds **in expectation**.

For this, we use that by definition,  $\mathbf{g}_t$  is an **unbiased estimate** of  $\nabla f(\mathbf{x}_t)$ :

$$\mathbb{E}[\mathbf{g}_t | \mathbf{x}_t = \mathbf{x}] = \frac{1}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}) = \nabla f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^d.$$

## The inequality $f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)$ holds in expectation

For any fixed  $\mathbf{x}$ , [linearity of conditional expectations](#) (Exercise 32) yields

$$\mathbb{E}[\mathbf{g}_t^\top (\mathbf{x} - \mathbf{x}^*) | \mathbf{x}_t = \mathbf{x}] = \mathbb{E}[\mathbf{g}_t | \mathbf{x}_t = \mathbf{x}]^\top (\mathbf{x} - \mathbf{x}^*) = \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^*).$$

Event  $\{\mathbf{x}_t = \mathbf{x}\}$  can occur only for  $\mathbf{x}$  in some finite set  $X$  ( $\mathbf{x}_t$  is determined by the choices of indices in all iterations so far). [Partition Theorem](#) (Exercise 32):

$$\begin{aligned} \mathbb{E}[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)] &= \sum_{\mathbf{x} \in X} \mathbb{E}[\mathbf{g}_t^\top (\mathbf{x} - \mathbf{x}^*) | \mathbf{x}_t = \mathbf{x}] \text{prob}(\mathbf{x}_t = \mathbf{x}) \\ &= \sum_{\mathbf{x} \in X} \nabla f(\mathbf{x})^\top (\mathbf{x} - \mathbf{x}^*) \text{prob}(\mathbf{x}_t = \mathbf{x}) = \mathbb{E}[\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*)]. \end{aligned}$$

Hence,

$$\mathbb{E}[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)] = \mathbb{E}[\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*)] \geq \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)].$$

$\downarrow$  convexity

## Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps

### Theorem

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be convex and differentiable,  $\mathbf{x}^*$  a global minimum; furthermore, suppose that  $\|\mathbf{x}_0 - \mathbf{x}^*\| \leq R$ , and that  $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq B^2$  for all  $t$ . Choosing the constant stepsize

$$\gamma := \frac{R}{B\sqrt{T}}$$

stochastic gradient descent yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(\mathbf{x}_t)] - f(\mathbf{x}^*) \leq \frac{RB}{\sqrt{T}}.$$

Same procedure as every week. . . except

- ▶ we assume bounded stochastic gradients **in expectation**;
- ▶ error bound holds **in expectation**.

## Bounded stochastic gradients: $\mathcal{O}(1/\varepsilon^2)$ steps II

Proof.

Vanilla analysis (this time,  $\mathbf{g}_t$  is the stochastic gradient):

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2.$$

Taking expectations and using “convexity in expectation”:

$$\begin{aligned} \sum_{t=0}^{T-1} \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] &\leq \sum_{t=0}^{T-1} \mathbb{E}[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)] \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \mathbb{E}[\|\mathbf{g}_t\|^2] + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^*\|^2 \\ &\leq \frac{\gamma}{2} B^2 T + \frac{1}{2\gamma} R^2. \end{aligned}$$

Result follows as every week (optimize  $\gamma$ ) ...





## Convergence rate comparison: SGD vs GD

**Classic GD:** For vanilla analysis, we assumed that  $\|\nabla f(\mathbf{x})\|^2 \leq B_{\text{GD}}^2$  for all  $\mathbf{x} \in \mathbb{R}^d$ , where  $B_{\text{GD}}$  was a constant. So for sum-objective:

$$\left\| \frac{1}{n} \sum_i \nabla f_i(\mathbf{x}) \right\|^2 \leq B_{\text{GD}}^2 \quad \forall \mathbf{x}$$

**SGD:** Assuming same for the **expected** squared norms of our stochastic gradients, now called  $B_{\text{SGD}}^2$ .

$$\frac{1}{n} \sum_i \|\nabla f_i(\mathbf{x})\|^2 \leq B_{\text{SGD}}^2 \quad \forall \mathbf{x}$$

So by Jensen's inequality for  $\|\cdot\|^2$

- ▶  $B_{\text{GD}}^2 \approx \left\| \frac{1}{n} \sum_i \nabla f_i(\mathbf{x}) \right\|^2 \leq \frac{1}{n} \sum_i \|\nabla f_i(\mathbf{x})\|^2 \approx B_{\text{SGD}}^2$
- ▶  $B_{\text{GD}}^2$  can be smaller than  $B_{\text{SGD}}^2$ , but often comparable. Very similar if larger mini-batches are used.

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps

### Theorem

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be differentiable and strongly convex with parameter  $\mu > 0$ ; let  $\mathbf{x}^*$  be the unique global minimum of  $f$ . With decreasing step size

$$\gamma_t := \frac{2}{\mu(t+1)}$$

stochastic gradient descent yields

$$\mathbb{E} \left[ f \left( \frac{2}{T(T+1)} \sum_{t=1}^T t \cdot \mathbf{x}_t \right) - f(\mathbf{x}^*) \right] \leq \frac{2B^2}{\mu(T+1)},$$

where  $B^2 := \max_{t=1}^T \mathbb{E} [\|\mathbf{g}_t\|^2]$ .

Almost same result as for subgradient descent, but **in expectation**.

## Tame strong convexity: $\mathcal{O}(1/\varepsilon)$ steps II

Proof.

Take expectations over vanilla analysis, **before** summing up (with varying stepsize  $\gamma_t$ ):

$$\mathbb{E}[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)] = \frac{\gamma_t}{2} \mathbb{E}[\|\mathbf{g}_t\|^2] + \frac{1}{2\gamma_t} (\mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2]).$$

“Strong convexity in expectation”:

$$\mathbb{E}[\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^*)] = \mathbb{E}[\nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^*)] \geq \mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] + \frac{\mu}{2} \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2]$$

Putting it together (with  $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq B^2$ ):

$$\mathbb{E}[f(\mathbf{x}_t) - f(\mathbf{x}^*)] \leq \frac{B^2\gamma_t}{2} + \frac{(\gamma_t^{-1} - \mu)}{2} \mathbb{E}[\|\mathbf{x}_t - \mathbf{x}^*\|^2] - \frac{\gamma_t^{-1}}{2} \mathbb{E}[\|\mathbf{x}_{t+1} - \mathbf{x}^*\|^2].$$

Proof continues as for subgradient descent, this time with expectations. □

# Mini-batch SGD

Instead of using a single element  $f_i$ , use an average of several of them:

$$\tilde{\mathbf{g}}_t := \frac{1}{m} \sum_{j=1}^m \mathbf{g}_t^j.$$

Extreme cases:

$m = 1 \Leftrightarrow$  SGD as originally defined

$m = n \Leftrightarrow$  full gradient descent

**Benefit:** Gradient computation can be naively parallelized

## Mini-batch SGD

**Variance Intuition:** Taking an average of many independent random variables reduces the variance. So for larger size of the mini-batch  $m$ ,  $\tilde{\mathbf{g}}_t$  will be closer to the true gradient, in expectation:

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\mathbf{g}}_t - \nabla f(\mathbf{x}_t)\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{1}{m} \sum_{j=1}^m \mathbf{g}_t^j - \nabla f(\mathbf{x}_t)\right\|^2\right] \\ &= \frac{1}{m} \mathbb{E}\left[\left\|\mathbf{g}_t^1 - \nabla f(\mathbf{x}_t)\right\|^2\right] \\ &= \frac{1}{m} \mathbb{E}\left[\left\|\mathbf{g}_t^1\right\|^2\right] - \frac{1}{m} \left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{B^2}{m} .\end{aligned}$$

Using a modification of the SGD analysis, can use this quantity to relate convergence rate to the rate of full gradient descent.

# Stochastic Subgradient Descent

For problems which are not necessarily differentiable, we modify SGD to use a subgradient of  $f_i$  in each iteration. The update of **stochastic subgradient descent** is given by

sample  $i \in [n]$  uniformly at random  
let  $\mathbf{g}_t \in \partial f_i(\mathbf{x}_t)$   
 $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma_t \mathbf{g}_t.$

In other words, we are using an **unbiased estimate of a subgradient** at each step,  $\mathbb{E}[\mathbf{g}_t | \mathbf{x}_t] \in \partial f(\mathbf{x}_t).$

Convergence in  $\mathcal{O}(1/\varepsilon^2)$ , by using the **subgradient property** at the beginning of the proof, where convexity was applied.

# Constrained optimization

For constrained optimization, our theorem for the SGD convergence in  $\mathcal{O}(1/\varepsilon^2)$  steps directly extends to constrained problems as well.

After every step of SGD, projection back to  $X$  is applied as usual. The resulting algorithm is called **projected SGD**.

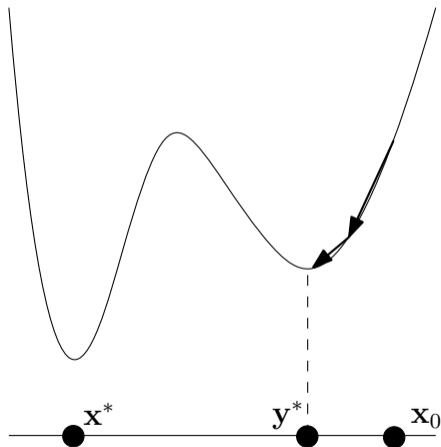
# Chapter 6

## Non-convex Optimization



# Gradient Descent in the nonconvex world

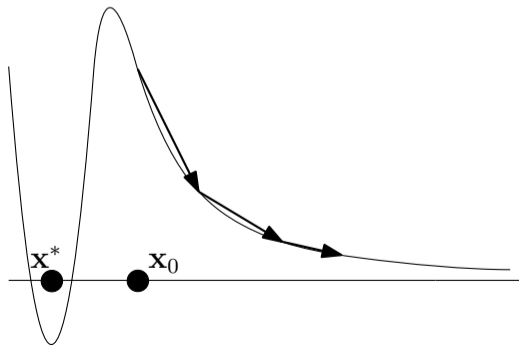
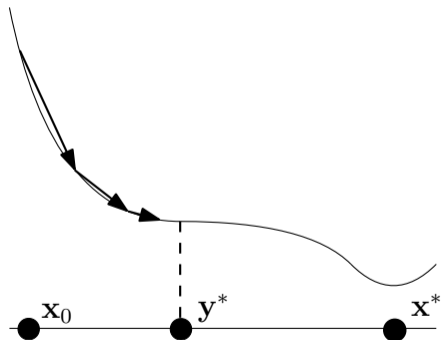
- ▶ may get stuck in a **local** minimum and miss the global minimum;



## Gradient Descent in the nonconvex world II

Even if there is a **unique** local minimum (equal to the global minimum), we

- ▶ may get stuck in a **saddle point**;
- ▶ run off to infinity;
- ▶ possibly encounter other bad behaviors.



# Gradient Descent in the nonconvex world III

Often, we observe good behavior in practice.

Theoretical explanations mostly missing.

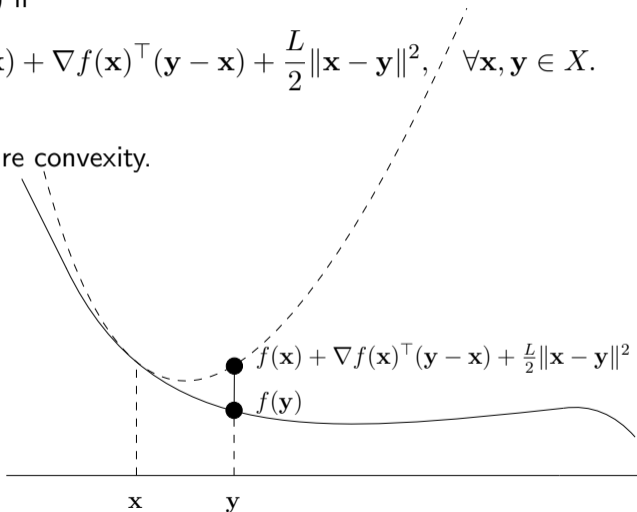
This lecture: under favorable conditions, we sometimes **can** say something useful about the behavior of gradient descent, even on nonconvex functions.

## Smooth (but not necessarily convex) functions

**Recall:** A differentiable  $f : \text{dom}(f) \rightarrow \mathbb{R}$  is smooth with parameter  $L \in \mathbb{R}_+$  over a convex set  $X \subseteq \text{dom}(f)$  if

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

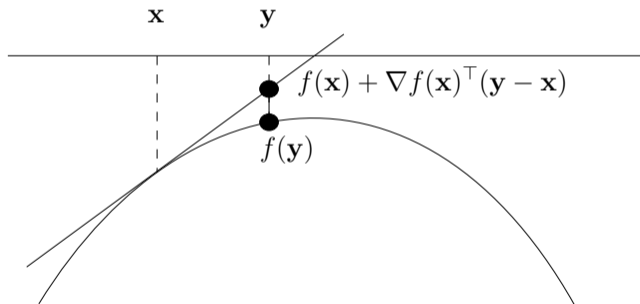
Definition does not require convexity.



## Concave functions

$f$  is called **concave** if  $-f$  is convex.

For all  $\mathbf{x}$ , the graph of a differentiable concave function is **below** the tangent hyperplane at  $\mathbf{x}$ .



$\Rightarrow$  concave functions are smooth with  $L = 0 \dots$  but boring from an optimization point of view (no global minimum), gradient descent runs off to infinity

## Bounded Hessians $\Rightarrow$ smooth

### Lemma

Let  $f : \mathbf{dom}(f) \rightarrow \mathbb{R}$  be twice differentiable, with  $X \subseteq \mathbf{dom}(f)$  a convex set, and  $\|\nabla^2 f(\mathbf{x})\| \leq L$  for all  $\mathbf{x} \in X$ , where  $\|\cdot\|$  is spectral norm. Then  $f$  is smooth with parameter  $L$  over  $X$ .

Examples:

- ▶ all quadratic functions  $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$
- ▶  $f(x) = \sin(x)$  (many global minima)

## Bounded Hessians $\Rightarrow$ smooth II

Proof.

By Theorem 1.10 (applied to the gradient function  $\nabla f$ ), bounded Hessians imply Lipschitz continuity of the gradient,

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

To show that this implies smoothness, we use  $h(1) - h(0) = \int_0^1 h'(t) dt$  with

$$h(t) := f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \quad t \in [0, 1],$$

Chain rule:

$$h'(t) = \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}).$$

## Bounded Hessians $\Rightarrow$ smooth III

Proof.

For  $\mathbf{x}, \mathbf{y} \in X$ :

$$\begin{aligned} & f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ = & h(1) - h(0) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \quad (\text{definition of } h) \\ = & \int_0^1 h'(t) dt - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ = & \int_0^1 \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) dt - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ = & \int_0^1 (\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})) dt \\ = & \int_0^1 (\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) dt \end{aligned}$$



## Bounded Hessians $\Rightarrow$ smooth IV

Proof.

For  $\mathbf{x}, \mathbf{y} \in X$ :

$$\begin{aligned} & f(\mathbf{y}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \\ = & \int_0^1 (\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) dt \\ \leq & \int_0^1 |(\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x})| dt \\ \leq & \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| dt \quad (\text{Cauchy-Schwarz}) \\ \leq & \int_0^1 L \|t(\mathbf{y} - \mathbf{x})\| \|\mathbf{y} - \mathbf{x}\| dt \quad (\text{Lipschitz continuous gradients (6.1)}) \\ = & \int_0^1 Lt \|\mathbf{x} - \mathbf{y}\|^2 = \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2. \end{aligned}$$

## Smooth $\Rightarrow$ bounded Hessians?

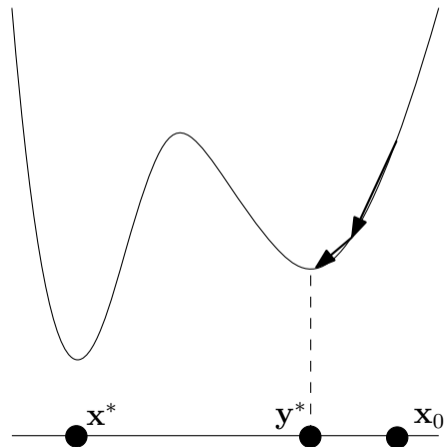
Yes, over any open convex set  $X$  (Exercise 33).

## Gradient descent on smooth functions

Will prove:  $\|\nabla f(\mathbf{x}_t)\|^2 \rightarrow 0$  for  $t \rightarrow \infty \dots$

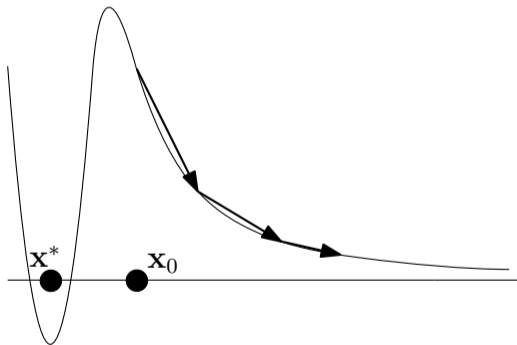
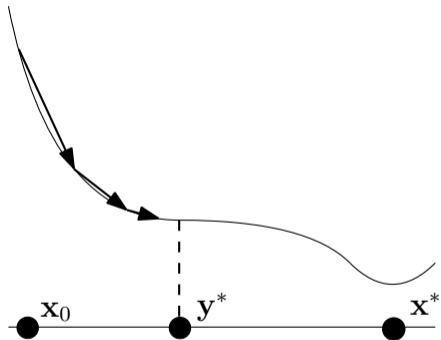
$\dots$  at the same rate as  $f(\mathbf{x}_t) - f(\mathbf{x}^*) \rightarrow 0$  in the convex case.

$f(\mathbf{x}_t) - f(\mathbf{x}^*)$  itself may **not** converge to 0 in the nonconvex case:



## What does $\|\nabla f(\mathbf{x}_t)\|^2 \rightarrow 0$ mean?

It may or **may not** mean that we converge to a **critical point** ( $\nabla f(\mathbf{y}^*) = \mathbf{0}$ )



# Gradient descent on smooth (not necessarily convex) functions

## Theorem

Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be differentiable with a global minimum  $\mathbf{x}^*$ ; furthermore, suppose that  $f$  is smooth with parameter  $L$  according to Definition 2.2. Choosing stepsize

$$\gamma := \frac{1}{L},$$

*gradient descent yields*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T} (f(\mathbf{x}_0) - f(\mathbf{x}^*)), \quad T > 0.$$

*In particular,  $\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T} (f(\mathbf{x}_0) - f(\mathbf{x}^*))$  for some  $t \in \{0, \dots, T-1\}$ .*

*And also,  $\lim_{t \rightarrow \infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0$  (Exercise 34).*

## Gradient descent on smooth (not necessarily convex) functions II

Proof.

Sufficient decrease (Lemma 2.6), does not require convexity:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

Rewriting:

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2L(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})).$$

Telescoping sum:

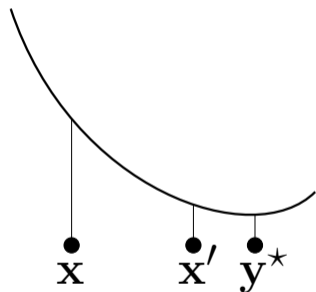
$$\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq 2L(f(\mathbf{x}_0) - f(\mathbf{x}_T)) \leq 2L(f(\mathbf{x}_0) - f(\mathbf{x}^*)).$$

The statement follows (divide by  $T$ ).

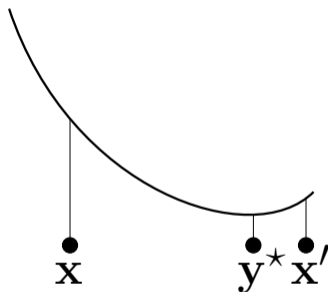


## No overshooting

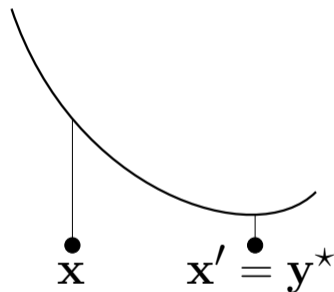
In the smooth setting, and with stepsize  $1/L$ , gradient descent cannot overshoot, i.e. pass a critical point (Exercise 35).



$$\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x}), \gamma < 1/L$$



overshooting



may happen with  $\gamma = 1/L$