Optimization for Machine Learning

Lecture 2b: Stochastic Gradient Descent and Non-convex optimization

Martin Jaggi
EPFL

PKU Summer School
`github.com/epfml/optml-pku`
August 1, 2023

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

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

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 $\nabla f(\mathbf{x}_t)$:

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

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

For any fixed $\mathbf{x}$, linearity of conditional expectations (Exercise 37) yields

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

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 37):

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

Hence,

$$\mathbb{E}\big[\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] = \mathbb{E}\big[\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t-\mathbf{x}^\star)\big] \overset{\downarrow\ \text{convexity}}{\geq} \mathbb{E}\big[f(\mathbf{x}_t)-f(\mathbf{x}^\star)\big].$$

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

Theorem

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

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

stochastic gradient descent yields

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

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):

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

Taking expectations and using "convexity in expectation":

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

Result follows as every week (optimize γ) ... $\square$

Convergence rate comparison: SGD vs GD

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

$$\left\| \frac{1}{n} \sum_i \nabla f_i(\mathbf{x}) \right\|^2 \le B_{\mathsf{GD}}^2 \qquad \forall \mathbf{x}$$

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

$$\frac{1}{n} \sum_i \left\| \nabla f_i(\mathbf{x}) \right\|^2 \le B_{\mathsf{SGD}}^2 \qquad \forall \mathbf{x}$$

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

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

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

Theorem

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

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

stochastic gradient descent yields

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

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

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 γ_t):

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

"Strong convexity in expectation":

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

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

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

Proof continues as for subgradient descent, this time with expectations. $\square$

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:

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

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, $\mathbb{E}\big[\mathbf{g}_t \big| \mathbf{x}_t\big] \in \partial f(\mathbf{x}_t)$.

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.

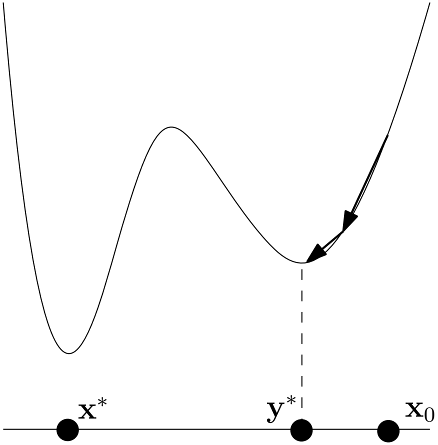
After every step of SGD, projection back to X is applied as usual. The resulting algorithm is called projected SGD.

Chapter 6

Non-convex Optimization

Gradient Descent in the nonconvex world

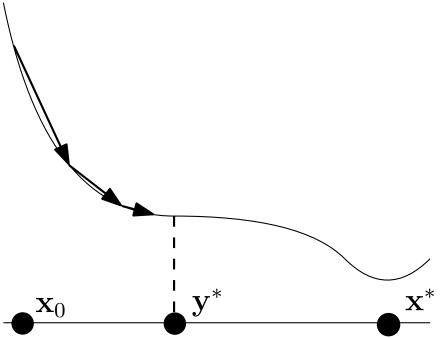
- may get stuck in a local minimum and miss the global minimum;

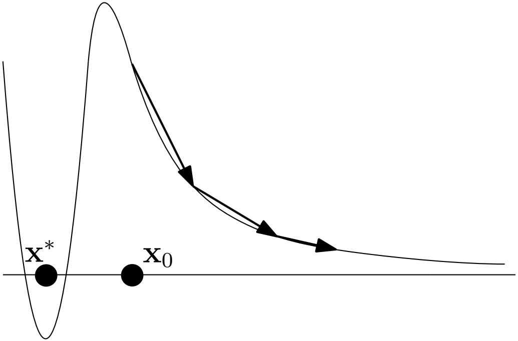


Gradient Descent in the nonconvex world II

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.

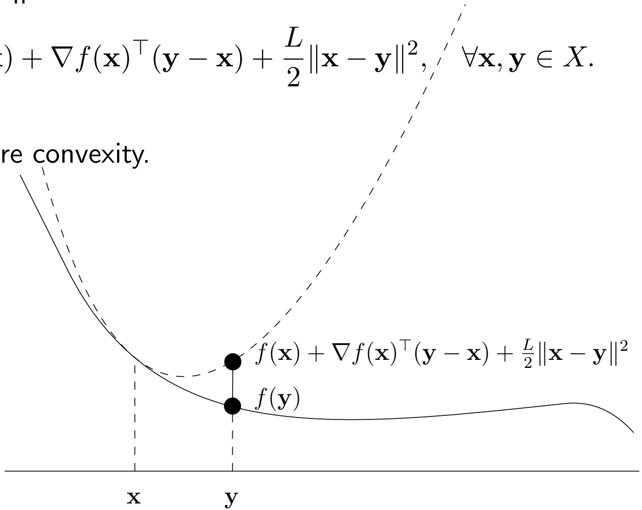
This lecture: under favorable conditions, we sometimes can say something useful about the behavior of gradient descent, even on nonconvex functions.

Smooth (but not necessarily convex) functions

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

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

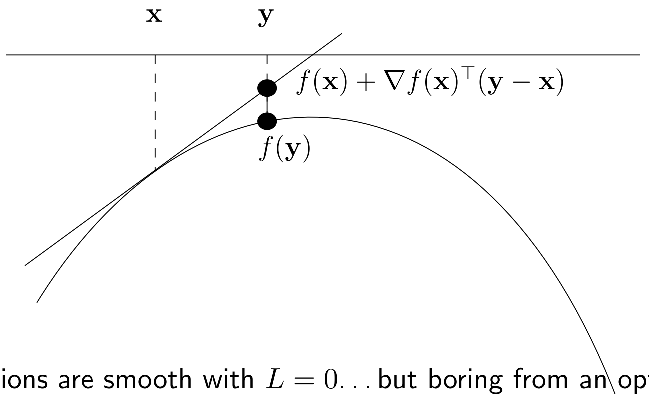
Definition does not require convexity.



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 $L = 0$... but boring from an optimization point of view (no global minimum), gradient descent runs off to infinity

Bounded Hessians $\Rightarrow$ smooth

Lemma

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be twice differentiable, with $X \subseteq \mathbf{dom}(f)$ a convex set, and $\left\|\nabla^2 f(\mathbf{x})\right\| \leq L$ 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 ∇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

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

Chain rule:

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

Bounded Hessians $\Rightarrow$ smooth III

Proof.

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

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

Bounded Hessians $\Rightarrow$ smooth IV

Proof.

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

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

□

Smooth $\Rightarrow$ bounded Hessians?

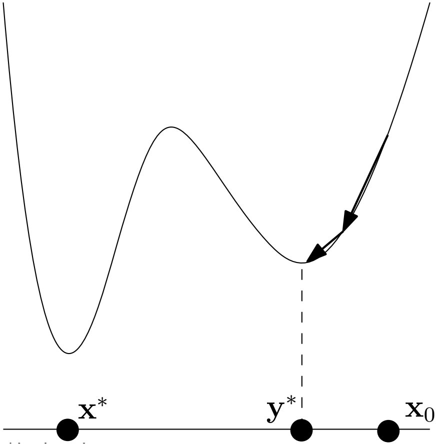
Yes, over any open convex set X (Exercise 38).

Gradient descent on smooth functions

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

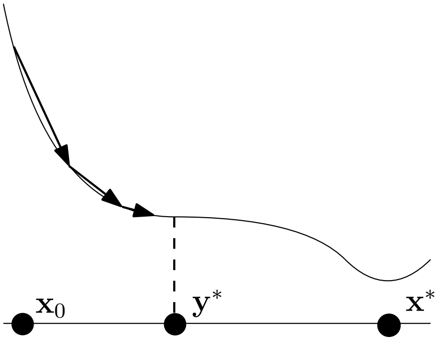
...at the same rate as $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \to 0$ in the convex case.

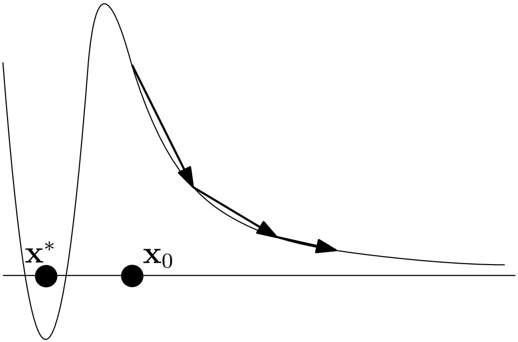
$f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ itself may not converge to 0 in the nonconvex case:



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

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





Gradient descent on smooth (not necessarily convex) functions

Theorem

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

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

gradient descent yields

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

In particular, $\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{2L}{T}\big(f(\mathbf{x}_0) - f(\mathbf{x}^\star)\big)$ for some $t \in \{0, \ldots, T-1\}$.
And also, $\lim_{t\to\infty} \|\nabla f(\mathbf{x}_t)\|^2 = 0$ (Exercise 39).

Gradient descent on smooth (not necessarily convex) functions II

Proof.

Sufficient decrease (Lemma 2.7), 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:

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

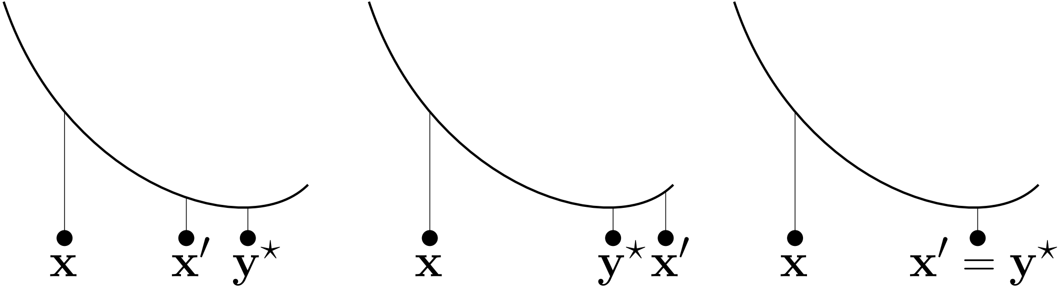
Telescoping sum:

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

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 40).



$\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x}), \gamma < 1/L$ | overshooting | may happen with $\gamma = 1/L$