Optimization for Machine Learning
CS-439

Lecture 2: Gradient Descent

Nicolas Flammarion

EPFL – github.com/epfml/OptML_course

March 3, 2022

Chapter 2

Gradient Descent

The Algorithm

Get near to a minimum $\mathbf{x}^\star$ / close to the optimal value $f(\mathbf{x}^\star)$?

(Assumptions: $f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, has a global minimum $\mathbf{x}^\star$)

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

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \varepsilon.$$

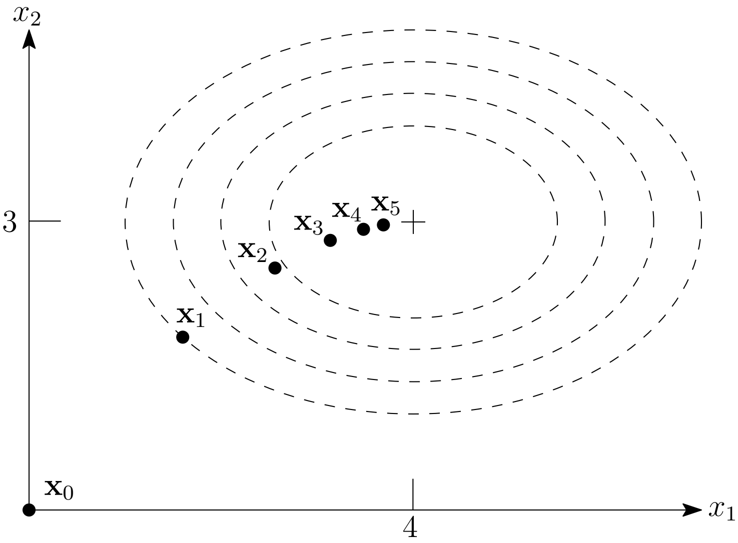
Note that there can be several global minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

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

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

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

Example



$$f(x_1, x_2) := 2(x_1 - 4)^2 + 3(x_2 - 3)^2, \mathbf{x}_0 := (0, 0), \gamma := 0.1$$

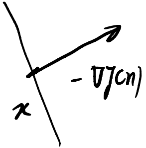
Intuition

locally $\boxed{f(x+v) \sim f(x) + \nabla f(x)^T v + O(\|v\|^2)}$

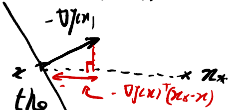
$\rightarrow$ v is a descent direction if $\nabla f(x)^T v \leq 0$

$\rightarrow$ How v? $\quad \min_{\|v\| \leq 1} v^T \nabla f(x) \quad \rightarrow \quad v = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$

- $\perp$ grad : fct varying like a square
- $\parallel$ grad : fct like a linear function



with convexity: $\quad f(x) - f(y) \leq \nabla f(x)^T (x-y) \quad \Rightarrow \quad \boxed{f(x) - f(x_*) \leq -\nabla f(x)^T (x_* - x)}$



$\Rightarrow$ rate of decrease of distance to the optimum is lowerbounded by the suboptimality gap

Continuous-time analysis

$$\frac{d}{dt} x(t) = -\nabla f(x(t)) \qquad \text{Gradient flow}$$

- Not implementable but only for the analysis
- $$\frac{d}{dt}\frac{1}{2}\|x(t) - x_*\|^2 = (x(t) - x_*)^T \dot{x}_t$$

$$= -(x(t) - x_*)^T \nabla f(x(t))$$

$$\overset{\text{cvx}}{\leq} -[f(x(t)) - f(x_*)]$$

$$\Rightarrow \quad f(x(t)) - f(x_*) \leq -\frac{d}{dt}\frac{1}{2}\|x(t) - x_*\|^2$$

$$\frac{1}{\tau}\int_0^{\tau}[f(x(t)) - f(x_*)]\,dt \leq \frac{1}{2\tau}\|x(0) - x_*\|^2 - \frac{1}{2\tau}\|x(\tau) - x_*\|^2$$

$$f\left(\frac{1}{\tau}\int_0^{\tau} x(t)\,dt\right) - f(x_*) \overset{\text{Jensen}}{\leq} \frac{1}{\tau}\int_0^{\tau}[f(x(t)) - f(x_*)]\,dt \leq \frac{1}{2\tau}\|x(0) - x_*\|^2$$

$\bar{x}_\tau = \frac{1}{\tau}\int_0^{\tau} x(t)\,dt$

$$\boxed{f(\bar{x}_\tau) - f(x_*) \leq \frac{1}{2\tau}\|x(0) - x_*\|^2}$$

$O(\frac{1}{\tau})$ rate!

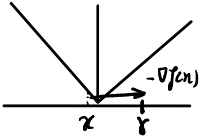
Intuition for the discretization error

GF is running GD with step size dt

After T steps GD will be close from GF at time $\tau = T\gamma$

$$x_T \sim x(T\gamma) \qquad f(\bar{x}_T) - f(x_*) \sim \frac{1}{2}\frac{\|x_0 - x_*\|^2}{T\gamma}$$

Ex:



$x \mapsto |x|$

$\Rightarrow$ need Lipschitz assumption on f
i.e. gradients are bounded

If $\|\nabla f(x)\| \le B$

$$f(\bar{x}_T) - f(x_*) \le \frac{1}{2}\frac{\|x_1 - x_*\|^2}{T\gamma} + \underbrace{\gamma B^2}_{\text{error from the discretizato}}$$

$$\le \frac{\|x_1 - x_*\| B}{\sqrt{T}} \quad \text{for } \gamma = \frac{\|x_1 - x_*\|}{B\sqrt{T}}$$

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$?

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top(\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top\mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ to rewrite

$$\begin{aligned}
\mathbf{g}_t^\top(\mathbf{x}_t-\mathbf{x}^\star) &= \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right) \\
&= \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t-\mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1}-\mathbf{x}^\star\|^2\right)
\end{aligned}$$

- Sum this up over the first T iterations:

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

Vanilla analysis II

Use first-order characterization of convexity: $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \forall \mathbf{x}, \mathbf{y}$

- with $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:
$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)$$
  giving
$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2,$$
  an upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ over the steps

- last iterate is not necessarily the best one
- stepsize is crucial

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

Assume that all gradients of f are bounded in norm.

- Equivalent to f being Lipschitz (Theorem 1.9; **Exercise 12**).
- Rules out many interesting functions (for example, the "supermodel" $f(x) = x^2$)

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$. Choosing the stepsize

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

gradient descent yields

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

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

Proof.

- Plug $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\mathbf{g}_t\| \leq B$ into Vanilla Analysis II:

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

- choose γ such that

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

  is minimized.
- Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{T}}$, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$.
- Dividing by T, the result follows.

$\square$

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

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

Advantages:

- dimension-independent (no d in the bound)!
- holds for both average, or best iterate

In Practice:
What if we don't know R and B? $\rightarrow$ **Exercise 16** (having to know R can't be avoided)

Smooth functions

"Not too curved"

Definition

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) over X if

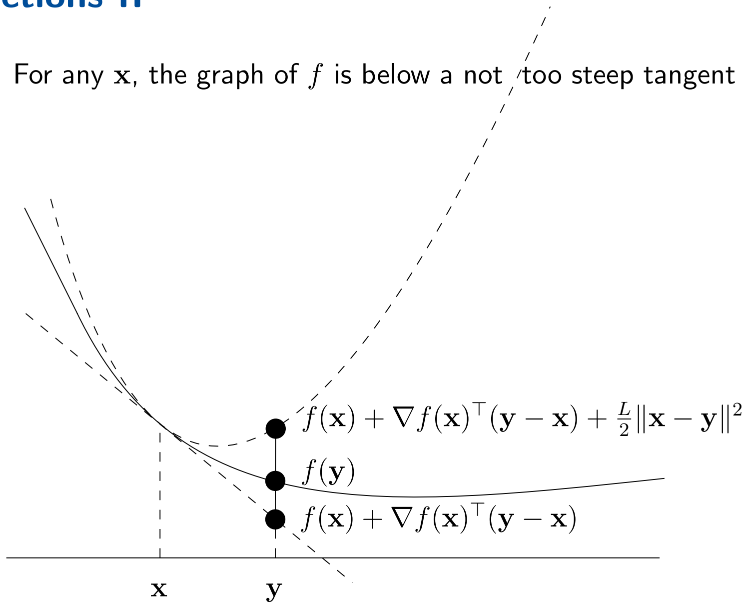
$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

f smooth $:\Leftrightarrow$ f smooth over $\mathbb{R}^d$.

Definition does not require convexity (useful later)

Smooth functions II

Smoothness: For any $\mathbf{x}$, the graph of f is below a not too steep tangent paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



Smooth functions III

- In general: quadratic functions are smooth (**Exercise 14**).
- Operations that preserve smoothness (the same that preserve convexity):

Lemma (Exercise 17)

(i) *Let $f_1, f_2, \ldots, f_m$ be 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 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 smooth with parameter L, and let $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for $A \in \mathbb{R}^{d \times m}$ and $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ is smooth with parameter $L\|A\|^2$, where is $\|A\|$ is the* ***spectral norm*** *of A (Definition 1.2).*

Smooth vs Lipschitz

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇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) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *for all* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

Proof in lecture slides of L. Vandenberghe, `http://www.seas.ucla.edu/~vandenbe/236C/lectures/gradient.pdf`.

Sufficient decrease

Lemma

Let $f : \mathbb{R}^d \to \mathbb{R}$ *be differentiable and smooth with parameter* L*. With stepsize*

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

gradient descent satisfies

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

Remark

More specifically, this already holds if f *is smooth with parameter* L *over the line segment connecting* $\mathbf{x}_t$ *and* $\mathbf{x}_{t+1}$*.*

Sufficient decrease II

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

Proof.

Use smoothness and definition of gradient descent ($\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$):

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

□

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 $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize

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

gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

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

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Vanilla Analysis II:

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

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

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

$\square$

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

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

$$
\begin{aligned}
\sum_{t=0}^{T-1} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) &\leq \frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\
&\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.
\end{aligned}
$$

Rewriting:

$$
\sum_{t=1}^{T} \left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) \leq \frac{L}{2} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.
$$

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

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

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

$R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$

$$T \geq \frac{R^2 L}{2\varepsilon} \quad \Rightarrow \quad \text{error } \leq \frac{L}{2T} R^2 \leq \varepsilon.$$

- $50 \cdot R^2 L$ 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?

$\rightarrow$ **Exercise 18**