Optimization for Machine Learning
CS-439

Lecture 2: Gradient Descent

Martin Jaggi

EPFL – `github.com/epfml/OptML_course`

March 2, 2018

Recap

Convexity

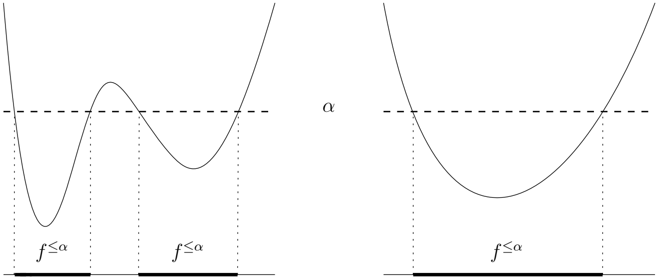
recap,
and short addition before we get to gradient descent...

Existence of a minimizer

Sublevel sets: Let $f : \mathbf{dom}(f) \to \mathbb{R}$, $\alpha \in \mathbb{R}$. The set

$$f^{\leq \alpha} := \{\mathbf{x} \in \mathbf{dom}(f) : f(\mathbf{x}) \leq \alpha\}$$

is the α-sublevel set of f;



Weierstrass Theorem

Theorem

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ *be a convex function,* $\mathbf{dom}(f)$ *open, and suppose there is a nonempty and bounded sublevel set* $f^{\leq \alpha}$*. Then* f *has a global minimum.*

Proof.

□

Chapter 2

Gradient Descent

The Algorithm

How to get near to a minimum $\mathbf{x}^\star$?

(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

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

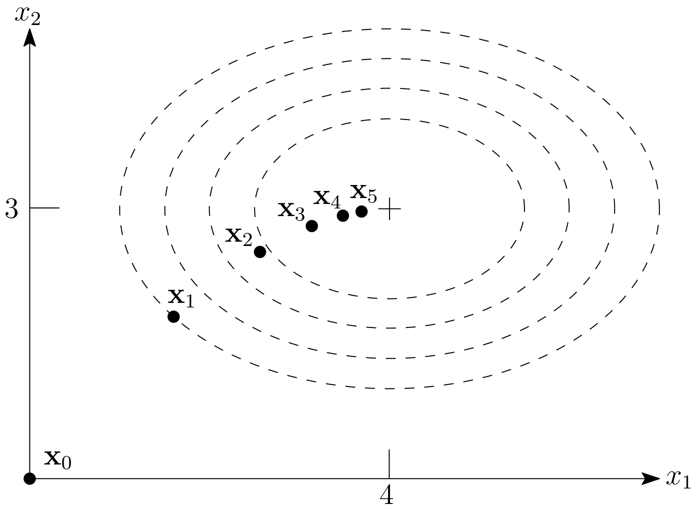
Note that there can be several minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

Iterative Algorithm:

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

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

Example


x_2
3
$\mathbf{x}_5$
$\mathbf{x}_4$
$\mathbf{x}_3$
$\mathbf{x}_2$
$\mathbf{x}_1$
$\mathbf{x}_0$
4
x_1

Vanilla analysis

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

- Convexity of f, for $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$, gives

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

- Apply the definition of the iteration, $\nabla f(\mathbf{x}_t) = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$:

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

- Now we apply $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$

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

again by the definition of gradient descent

Vanilla analysis, cont.

sum this over steps $t = 0, \dots, T-1$:

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

an upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$, $t = 0 \dots T-1$

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

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

Assume that all gradients of f are bounded in norm.

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* $\|\nabla f(\mathbf{x})\| \leq L$ *for all* $\mathbf{x}$*. Choosing the stepsize*

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

gradient descent yields

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

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

Proof.

□

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

Advantages:

- dimension-independent!
- holds for both average, or best iterate

In Practice:

What if we don't know R and L?

→ Exercise 12

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

Convex, but not too convex?

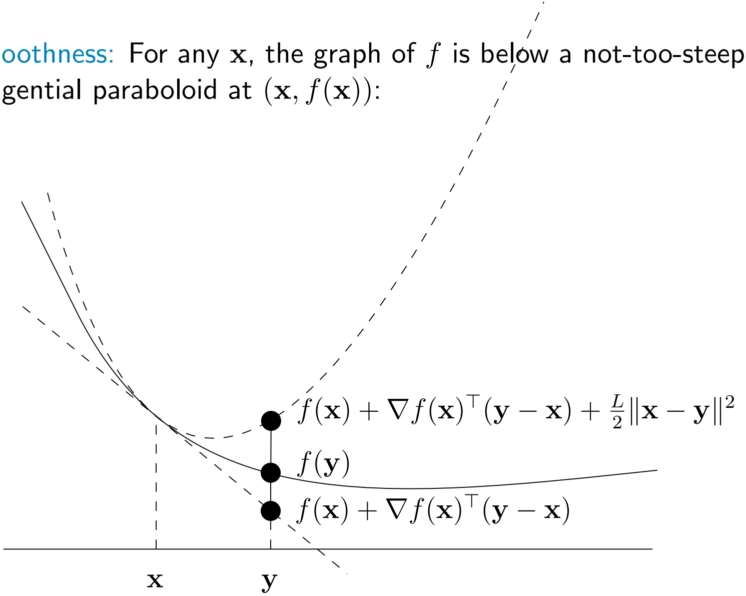
Definition

Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be convex and differentiable, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) if

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

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

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



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

- Quadratic functions are smooth
- Operations that preserve smoothness:

Lemma (Exercise 14)

(i) *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* $\sum_{i=1}^m \lambda_i L_i$*.*

(ii) *Let* f *be convex and 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 convex function* $f \circ g$ *is smooth with parameter* $L\|A\|^2$*, where*

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

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

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

Convergence proof: See next lecture