annotated version

# Optimization for Machine Learning
# CS-439

## Lecture 2: Gradient Descent

**Martin Jaggi**

EPFL – `github.com/epfml/OptML_course`

March 5, 2021

# Chapter 2

## Gradient Descent

github.com/opt_ml_course

# 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

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

accuracy

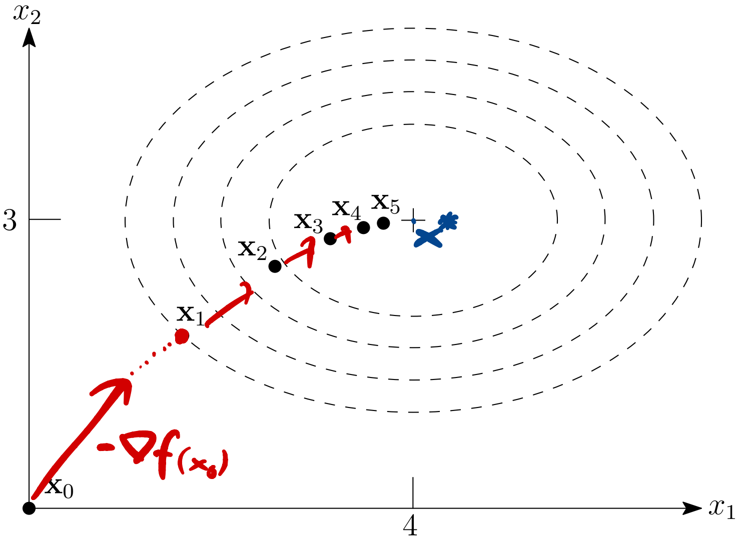
Note that there can be several global minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

**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** $\gamma \ge 0$.

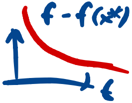
# Example




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

## Vanilla analysis

current error




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

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

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

def of step

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

identity

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

- Sum this up over the first $T$ iterations:

telescoping

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

# 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 $\mathbf{x} = \mathbf{x}_t, \mathbf{y} = \mathbf{x}^\star$:

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

prev. slide

giving

$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad \cancel{-\frac{1}{2\gamma}\|x_T - x^\star\|^2}$$

non-positive

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*

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

converge $\mathcal{O}(1/\sqrt{T})$

# 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:

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

- choose $\gamma$ such that

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

is minimized.

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

□

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

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

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 15** (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

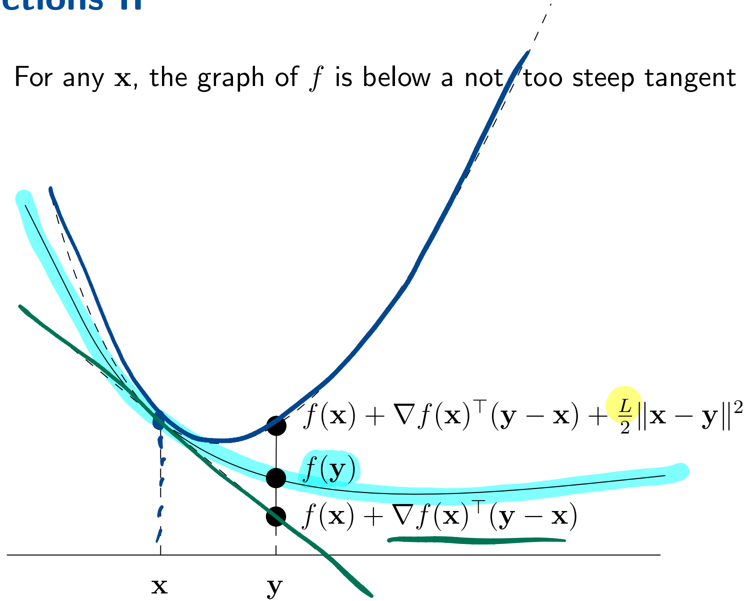
$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

$f$ 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 13**).
- Operations that preserve smoothness (the same that preserve convexity):

**Lemma (Exercise 16)**

(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 $\nabla f$ (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) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ *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*

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \underbrace{\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2}_{\text{improvement} > 0}, \quad t \ge 0.$$

## 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

$x_{t+1} - x_t = -\frac{1}{L}\nabla f(x_t)$

Lemma

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

Proof.

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

Smoothness

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

$\square$

# 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*

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

last iterate

# 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:

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

sufficient decrease

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

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

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

sufficient decrease

Rewriting:

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

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

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

$\square$

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

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

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

- $50 \cdot R^2 L$ iterations for error $0.01 = \varepsilon$ ...
- ... as opposed to $\underbrace{10,000}_{\leftarrow 1/\varepsilon^2} \cdot R^2 B^2$ in the Lipschitz case

much faster than

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

$\rightarrow$ **Exercise 17**