annotated version

Optimization for Machine Learning
CS-439

Lecture 3: Faster, and Projected Gradient Descent

Martin Jaggi

EPFL – github.com/epfml/OptML_course

March 9, 2018

Smooth 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

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

gradient descent with arbitrary $\mathbf{x}_0$ satisfies

(i) *Function values are monotone decreasing:*

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

(ii)

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

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

Proof. Smoothness

$-\frac{1}{L}\nabla f(x)$ by step definition

$$f(y) \leq f(x) + \nabla_f^T \underbrace{(y-x)}_{= -\frac{1}{L}\nabla f(x)} + \frac{L}{2}\|y-x\|^2$$

$$= f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2$$

$$= f(x) - \frac{1}{2L}\|.\|^2$$

$$f(y_t) - f(x_t) \leq -\frac{1}{2L}\|\nabla f\|^2 \quad \forall t$$

$$\circledast \quad \frac{1}{2L}\sum_{t=0}^{T}\|\nabla f(x_t)\|^2 \leq f(x_0) - f(x_T)$$

t

monotone decreasing $\Rightarrow f_T - f^* \leq \text{avg}(f_t) \leq \frac{L}{2T}\|..\|^2$ $\square$

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

- Do we need to know L?
  No. Exercise 15.

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

slow

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of f, $O(1/\varepsilon^2)$
- Now: smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f. $O(1/\varepsilon)$ fast

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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.*

Can we go even faster?

So far: Error decreases with $1/\sqrt{T}$, or $1/T$...

Could it decrease exponentially in T?

Can we go even faster?

- On $f(x) := x^2$: Stepsize $\gamma := \frac{1}{2}$ (f is $L = 2$ - smooth)

$$x_{t+1} = x_t - \frac{1}{2}\nabla f(x_t) = x_t - x_t = 0,$$

  - converged in one step!

- Same $f(x) := x^2$: Stepsize $\gamma := \frac{1}{4}$ (f is $L = 4$ - smooth)

$$x_{t+1} = x_t - \frac{1}{4}\nabla f(x_t) = x_t - \frac{x_t}{2} = \frac{x_t}{2},$$

  so $f(x_t) = f\left(\frac{x_0}{2^t}\right) = \frac{1}{2^{2t}} x_0^2$.

  - Exponential in t !

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps

Not too curved and not too flat

Definition

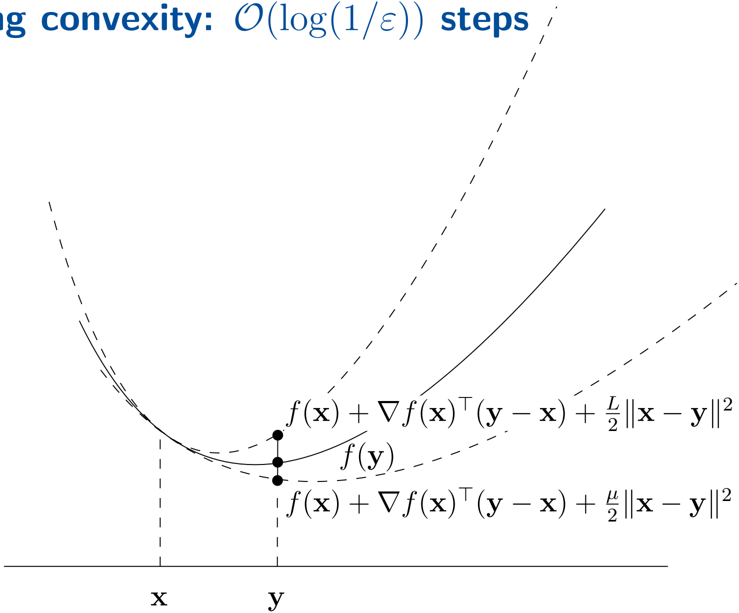
Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $\mu \in \mathbb{R}_+, \mu > 0$. f is called strongly convex (with parameter μ) if

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

Lemma (Exercise 17)

If f is strongly convex with parameter $\mu > 0$, then f is strictly convex and has a unique global minimum.

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps



A smooth and strongly convex function

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps

Can we show $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^\star$?

From the vanilla analysis, we know

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right).$$

Using that f is strongly convex, we obtain

$$\le \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \boxed{\|\mathbf{x}_t - \mathbf{x}^\star\|^2} - \boxed{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2}\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Can bound $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2$ in terms of $\|\mathbf{x}_t - \mathbf{x}^\star\|^2$, along with some "noise":

$$\boxed{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2} \le \underbrace{2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2}_{\text{noise}} + \underbrace{(1-\mu\gamma)}_{=\left(1-\frac{\mu}{L}\right)}\boxed{\|\mathbf{x}_t - \mathbf{x}^\star\|^2} \quad \text{(S)}$$

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and smooth with parameter L, and strongly convex with parameter $\mu > 0$. Choosing

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

gradient descent with arbitrary $\mathbf{x}_0$ satisfies the following two properties.

(i) *Squared distances to $\mathbf{x}^\star$ are geometrically decreasing:*

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

(ii)

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

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps

Proof.

For (i), we show that the noise in (S) disappears. From the above "smooth" Theorem (i), we know that

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

and hence the noise can be bounded as follows:

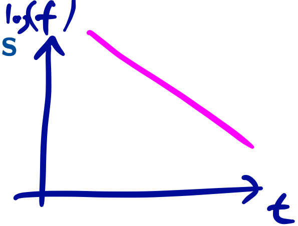
$$\begin{aligned}
& \boxed{2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2} \\
= \;& \frac{2}{L}(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 \\
\leq \;& -\frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2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

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

So, (S) actually yields

$$\boxed{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2} \leq (1-\mu\gamma)\boxed{\|\mathbf{x}_t - \mathbf{x}^\star\|^2} = \left(1 - \frac{\mu}{L}\right)\boxed{\|\mathbf{x}_t - \mathbf{x}^\star\|^2.}$$

Strong convexity: $\mathcal{O}(\log(1/\varepsilon))$ steps



Proof.

The bound in (ii) follows from smoothness, using $\nabla f(\mathbf{x}^\star) = \mathbf{0}$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \underbrace{\cancel{\nabla f(\mathbf{x}^\star)^\top(\mathbf{x}_t - \mathbf{x}^\star)}}_{=0} + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_t\|^2 = \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_t\|^2.$$

□

Conclusion: To reach absolute error at most ε, we only need $\mathcal{O}(\log \frac{1}{\varepsilon})$ iterations, where the constant behind the big-$\mathcal{O}$ is roughly L/μ.

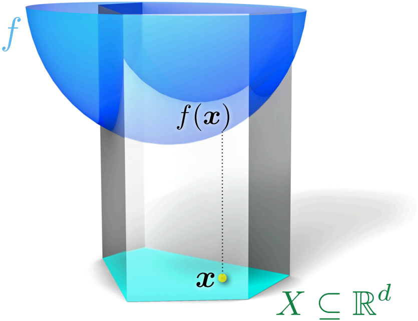
Chapter 3

Projected Gradient Descent

Constrained Optimization

Constrained Optimization Problem

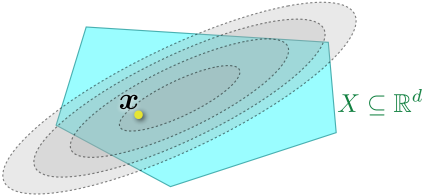
$$
\begin{array}{ll}
\text{minimize} & f(\mathbf{x}) \\
\text{subject to} & \mathbf{x} \in X
\end{array}
$$



Solving Constrained Optimization Problems

A Projected Gradient Descent

B Transform it into an *unconstrained* problem



The Algorithm

How to get near to a minimum $\mathbf{x}^\star$ over a closed convex subset $X \subseteq \mathbb{R}^d$?

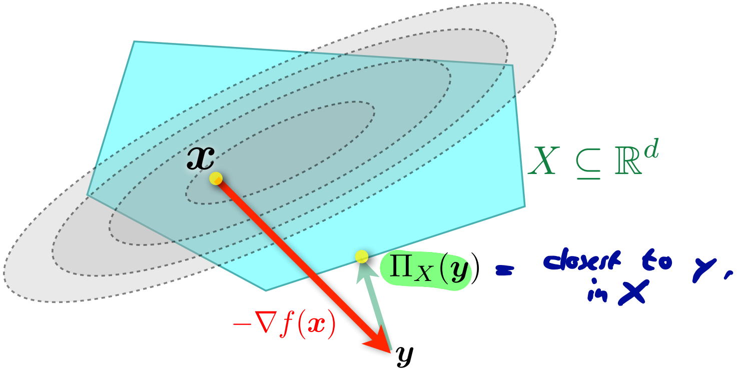
Projected gradient descent:

$$
\begin{aligned}
\mathbf{y}_{t+1} &:= \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t), \\
\mathbf{x}_{t+1} &:= \Pi_X(\mathbf{y}_{t+1}) := \underset{\mathbf{x} \in X}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{y}_{t+1}\|^2.
\end{aligned}
$$

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

Projected Gradient Descent

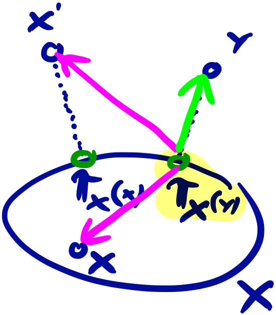
Idea: project onto X after every step:
$\Pi_X(\mathbf{y}) := \text{argmin}_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|$



Projected gradient update $\mathbf{x}_{t+1} \leftarrow \Pi_X\left[\mathbf{x}_t - \gamma\nabla f(\mathbf{x}_t)\right]$

Properties of Projection

contractivity



Fact

Let $X \subseteq \mathbb{R}^d$ convex, $\mathbf{x} \in X, \mathbf{y} \in \mathbb{R}^d$. Then

(i) $(\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y})) \leq 0$.

(ii) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

optimize

$$y \rightarrow \Pi_X(y) = \operatorname*{argmin}_{x \in X} \|x - y\|^2$$

Exercise !

Constrained minimization: $\mathcal{O}(1/\varepsilon^2)$ steps

„slow"

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable, $X \subseteq \mathbb{R}^d$ closed and convex, $\mathbf{x}^\star$ a minimizer of f over X; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ with $\mathbf{x}_0 \in X$, and that $\|\nabla f(\mathbf{x})\| \leq L$ for all $\mathbf{x} \in X$. Choosing the constant stepsize

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

projected gradient descent yields

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

Constrained minimization: $\mathcal{O}(1/\varepsilon^2)$ steps

Proof.

Vanilla analysis, but in early step, replace $\mathbf{x}_{t+1}$ by $\mathbf{y}_{t+1}$:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma} \left(\gamma^2 \|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2\right). \tag{1}$$

From Fact(ii) (with $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$), we obtain $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{y}_{t+1} - \mathbf{x}^\star\|^2$, hence we get

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

and follow the vanilla analysis for the remainder of the proof. □