Optimization for Machine Learning
CS-439

Lecture 3: Faster, and Projected Gradient Descent

Nicolas Flammarion

EPFL – github.com/epfml/OptML_course

March 7, 2025

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 !

Strongly convex functions

"Not too flat"

Definition

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subseteq \mathbf{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called strongly convex (with parameter μ) over X if

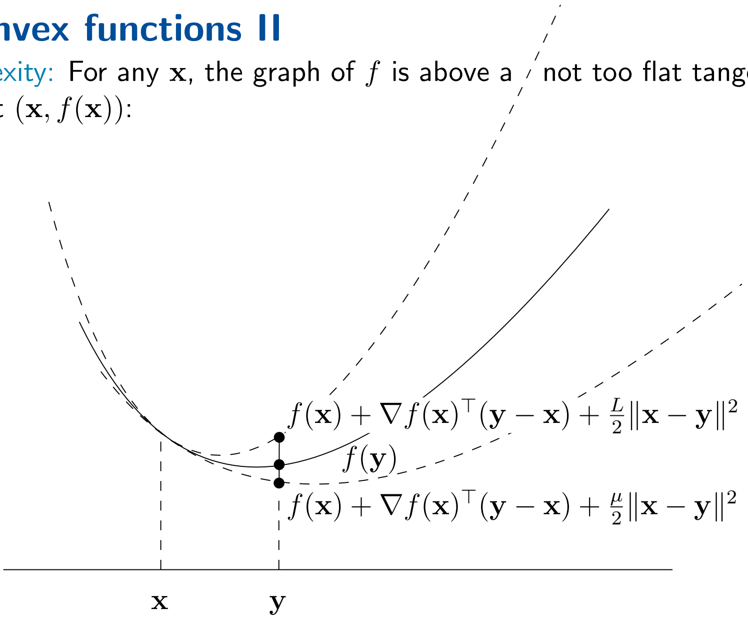
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

Lemma (Exercise 21)

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

Strongly convex functions II

Strong convexity: For any $\mathbf{x}$, the graph of f is above a not too flat tangential paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



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

Want to show: $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^\star$

Vanilla Analysis:

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

Now use stronger lower bound on left hand side, coming from strong convexity:

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

Putting it together:

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

Rewriting:

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

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

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

Squared distance to $\mathbf{x}^\star$ goes down by a constant factor, up to some "noise".

Theorem

Let $f:\mathbb{R}^d\to\mathbb{R}$ be differentiable with a global minimum $\mathbf{x}^\star$; suppose that f is 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:*

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

(ii) *The absolute error after T iterations is exponentially small in T:*

$$f(\mathbf{x}_T)-f(\mathbf{x}^\star) \le \frac{L}{2}\left(1-\frac{\mu}{L}\right)^T\|\mathbf{x}_0-\mathbf{x}^\star\|^2, \quad T>0.$$

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

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

Proof of (i).

Bounding the noise: $\gamma = 1/L$, sufficient decrease

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

Hence, the noise is nonpositive, and we get (i):

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

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

Proof of (ii).

From (i):

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

Smoothness together with $\nabla f(\mathbf{x}^\star) = \mathbf{0}$:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}^\star)^\top (\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 = \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2.$$

Putting it together:

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

$\square$

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

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

$$T \geq \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right) \quad \Rightarrow \quad \text{error } \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \leq \varepsilon.$$

Conclusion: To reach absolute error at most ε, we only need $\mathcal{O}(\log \frac{1}{\varepsilon})$ iterations, e.g.

- $\frac{L}{\mu} \ln(50 \cdot R^2 L)$ iterations for error 0.01 ...
- ... as opposed to $50 \cdot R^2 L$ in the smooth case

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

$\rightarrow$ (similar to) **Exercise 15**

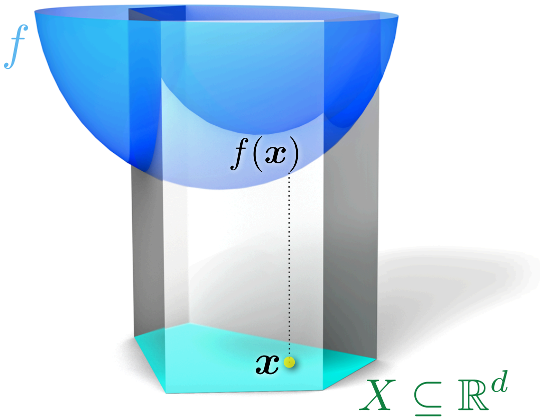
Chapter 3

Projected Gradient Descent

Constrained Optimization

Constrained Optimization Problem

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



Solving Constrained Optimization Problems

A Projected Gradient Descent

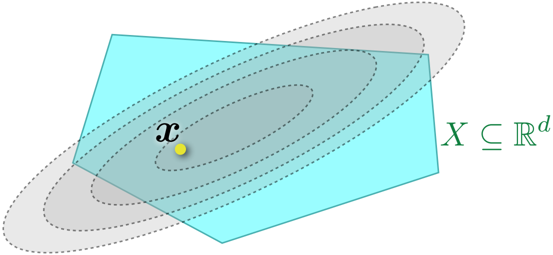
B Transform it into an *unconstrained* problem

Constrained Optimization

Solving Constrained Optimization Problems

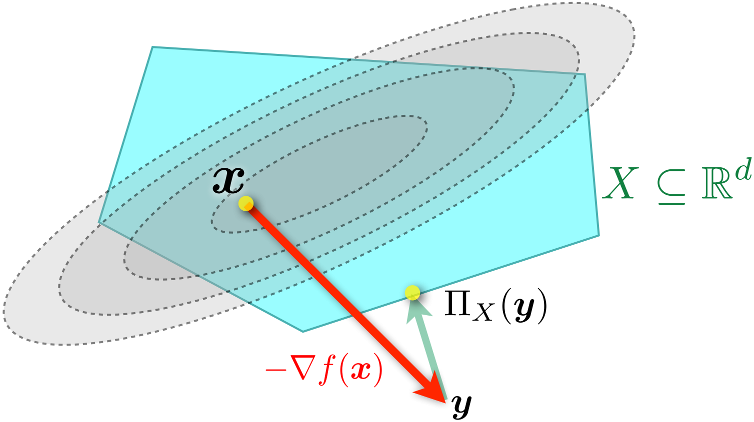
$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in X \end{array}$$

- Here: Projected Gradient Descent



Projected Gradient Descent

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



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

The Algorithm

Projected gradient descent:

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

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

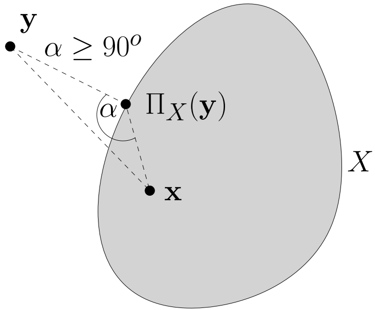
Properties of Projection

Fact

Let $X \subseteq \mathbb{R}^d$ be closed and 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) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.



Properties of Projection II

Fact

Let $X \subseteq \mathbb{R}^d$ be closed and 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) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Proof.

(i) $\Pi_X(\mathbf{y})$ is minimizer of (differentiable) convex function $d_{\mathbf{y}}(\mathbf{x}) = \|\mathbf{x} - \mathbf{y}\|^2$ over X. By first-order characterization of optimality (**Lemma 1.28**),

$$
\begin{array}{rcl}
 & 0 \;\leq & \nabla d_{\mathbf{y}}(\Pi_X(\mathbf{y}))^\top (\mathbf{x} - \Pi_X(\mathbf{y})) \\
 & = & 2(\Pi_X(\mathbf{y}) - \mathbf{y})^\top (\mathbf{x} - \Pi_X(\mathbf{y})) \\
\Leftrightarrow & 0 \;\geq & 2(\mathbf{y} - \Pi_X(\mathbf{y}))^\top (\mathbf{x} - \Pi_X(\mathbf{y})) \\
\Leftrightarrow & 0 \;\geq & (\mathbf{x} - \Pi_X(\mathbf{y}))^\top (\mathbf{y} - \Pi_X(\mathbf{y}))
\end{array}
$$

□

Properties of Projection III

Fact

Let $X \subseteq \mathbb{R}^d$ be closed and 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) $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.

Proof.

(ii)

$$\mathbf{v} := (\mathbf{x} - \Pi_X(\mathbf{y})), \quad \mathbf{w} := (\mathbf{y} - \Pi_X(\mathbf{y})).$$

By (i),

$$\begin{aligned} 0 \geq 2\mathbf{v}^\top \mathbf{w} &= \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2 \\ &= \|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 - \|\mathbf{x} - \mathbf{y}\|^2. \end{aligned}$$

$\square$

Results for projected gradient descent over closed and convex X

The same number of steps as gradient over $\mathbb{R}^d$!

- Lipschitz convex functions over X: $\mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions over X: $\mathcal{O}(1/\varepsilon)$ steps
- Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps

We will adapt the previous proofs for gradient descent.

BUT:

- Each step involves a projection onto X
- may or may not be efficient (in relevant cases, it is)...

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

Assume that all gradients of f are bounded in norm over closed and convex X.

- Equivalent to f being Lipschitz over X (Theorem 1.10; Exercise 12).
- Many interesting functions are Lipschitz over bounded sets X.

Theorem (same as the unconstrained one, but more useful)

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 $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x} \in X$. Choosing the constant stepsize

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

projected 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.

- Replace $\mathbf{x}_{t+1}$ in the vanilla analysis with $\mathbf{y}_{t+1}$ (the unprojected gradient step):

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\underline{\mathbf{y}_{t+1}} - \mathbf{x}^\star\|^2\right).$$

- Use Fact (ii): $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.
- With $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$, we have $\Pi_X(\mathbf{y}) = \mathbf{x}_{t+1}$, and hence

$$\|\mathbf{x}^\star - \mathbf{x}_{t+1}\|^2 \leq \|\mathbf{x}^\star - \mathbf{y}_{t+1}\|^2$$

- We go back to the original vanilla analyis and continue from there as before:

$$\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\underline{\mathbf{x}_{t+1}} - \mathbf{x}^\star\|^2\right).$$

□

Smooth functions over X

Recall:

f is called smooth (with parameter L) over X if

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

Sufficient decrease

Lemma

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

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

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

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

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

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

Proof.

Use smoothness, $\mathbf{y}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$, $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v} - \mathbf{w}\|^2$:

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

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

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a closed convex set, and assume that there is a minimizer $\mathbf{x}^\star$ of f over X; furthermore, suppose that f is smooth over X with parameter L. Choosing stepsize

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

projected 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 over X: $\mathcal{O}(1/\varepsilon)$ 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.

As before, use sufficient decrease to bound sum of squared gradients in vanilla analysis:

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

But now: extra term $\frac{L}{2}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$.

Compensate in the vanilla analysis itself! □

Recall: Constrained vanilla analysis

Proof.

- Replace $\mathbf{x}_{t+1}$ in the vanilla analysis with $\mathbf{y}_{t+1}$ (the unprojected gradient step):

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

- Use Fact (ii): $\|\mathbf{x} - \Pi_X(\mathbf{y})\|^2 + \|\mathbf{y} - \Pi_X(\mathbf{y})\|^2 \leq \|\mathbf{x} - \mathbf{y}\|^2$.
  With $\mathbf{x} = \mathbf{x}^\star, \mathbf{y} = \mathbf{y}_{t+1}$, we have $\Pi_X(\mathbf{y}) = \mathbf{x}_{t+1}$, and hence

$$\|\mathbf{x}^\star - \mathbf{x}_{t+1}\|^2 + \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 \quad \leq \|\mathbf{x}^\star - \mathbf{y}_{t+1}\|^2$$

- We get back to the vanilla analysis... but with a saving!

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

□

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

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

Proof.

Use $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star)$ (convexity), vanilla analysis with saving, $\gamma = 1/L$:

$$\begin{aligned}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \sum_{t=0}^{T-1}\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) \\ &\leq \frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \underline{\frac{L}{2}\sum_{t=0}^{T-1}\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2}.\end{aligned}$$

Use sufficient decrease to bound $\frac{1}{2L}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2$ by

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

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

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

Proof.

Putting it together: extra terms cancel, and as in unconstrained case, we get

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

Exercise 24: again, we make progress in every step (not immediate from sufficient decrease here). Hence,

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

□

Smooth and strongly convex functions over X

Recall:

f is strongly convex (with parameter μ) over X if

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

Smooth and strongly convex functions over X

Exercise 25: a strongly convex function has a unique minimizer $\mathbf{x}^\star$ of f over X.

We prove that projected gradient descent converges to $\mathbf{x}^\star$.

Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Let $X \subseteq \mathbb{R}^d$ be a nonempty closed and convex set and suppose that f is smooth over X with parameter L and strongly convex over X with parameter $\mu > 0$. Choosing $\gamma := \frac{1}{L}$, projected 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) *The absolute error after T iterations is exponentially small in T:*

$$\begin{aligned} f(\mathbf{x}_T) - f(\mathbf{x}^\star) &\leq \|\nabla f(\mathbf{x}^\star)\| \left(1 - \frac{\mu}{L}\right)^{T/2} \|\mathbf{x}_0 - \mathbf{x}^\star\| && \leftarrow \textit{in general, } \nabla f(\mathbf{x}^\star) \neq \mathbf{0}\textit{!} \\ &+ \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{T} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0. && \leftarrow \textit{as in unconstrained case} \end{aligned}$$

Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps I

Proof.

(i) Geometric decrease plus noise: $\underline{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2} \leq \cdots$

- unconstrained case:

$$2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 \qquad\qquad\qquad\qquad + \underline{(1-\mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2}.$$

- constrained case (vanilla analysis with a saving):

$$2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 - \|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2 + \underline{(1-\mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2}.$$

Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps II

Proof.

To bound the noise, we use sufficient decrease.

- unconstrained case:

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

- constrained case:

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

Putting it together, the terms $\|\mathbf{y}_{t+1} - \mathbf{x}_{t+1}\|^2$ cancel, and we get

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

in both cases. □

Smooth and strongly convex functions over X: $\mathcal{O}(\log(1/\varepsilon))$ steps III

Proof.

(ii) Error bound from smoothness:

$$\begin{aligned}
f(\mathbf{x}_T) - f(\mathbf{x}^\star) &\leq \nabla f(\mathbf{x}^\star)^\top(\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_T\|^2 \\
&\leq \|\nabla f(\mathbf{x}^\star)\| \, \|\mathbf{x}_T - \mathbf{x}^\star\| + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{x}_T\|^2 \text{ (Cauchy-Schwarz)} \\
&\leq \|\nabla f(\mathbf{x}^\star)\| \left(1 - \frac{\mu}{L}\right)^{T/2} \|\mathbf{x}_0 - \mathbf{x}^\star\| + \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^{T} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2 . \text{ (i)}
\end{aligned}$$

□

constrained error bound $\approx \sqrt{\text{unconstrained error bound}}$

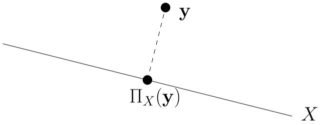
required number of steps roughly doubles.

The Projection Step: $\Pi_X(\mathbf{y}) := \operatorname{argmin}_{\mathbf{x}\in X} \|\mathbf{x}-\mathbf{y}\|$

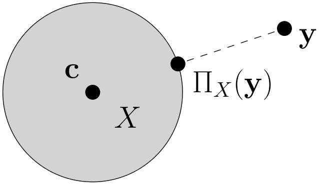
Computing $\Pi_X(\mathbf{y})$ is an optimization problem itself.

It can efficiently be solved in relevant cases:

- Projecting onto an affine subspace (leads to system of linear equations, similar to least squares)

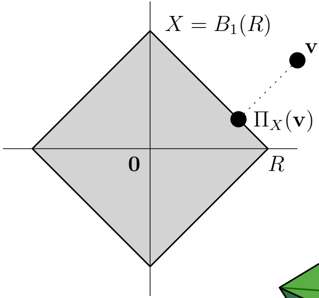


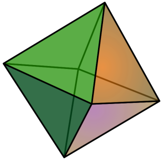
- Projecting onto a Euclidean ball with center $\mathbf{c}$ (simply scale the vector $\mathbf{y}-\mathbf{c}$)



Projecting onto ℓ_1-balls (needed in Lasso)

W.l.o.g. restrict to center at $\mathbf{0}$: $B_1(R) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_1 = \sum_{i=1}^d |x_i| \leq R\}$.



$B_1(R)$ is the cross polytope ($2d$ vertices, 2^d facets).  (octahedron, $d = 3$)

Section 3.5: projection can be computed in $\mathcal{O}(d \log d)$ time (can be improved to $\mathcal{O}(d)$)