

# Conditional gradient methods. Projected Gradient Descent. Frank-Wolfe Method. Mirror Descent Algorithm Idea.

Seminar

Optimization for ML. Faculty of Computer Science. HSE University

# Projection

The **distance**  $d$  from point  $\mathbf{y} \in \mathbb{R}^n$  to closed set  $S \subset \mathbb{R}^n$ :

$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|x - \mathbf{y}\| \mid x \in S\}$$

We will focus on **Euclidean projection** (other options are possible) of a point  $\mathbf{y} \in \mathbb{R}^n$  on set  $S \subseteq \mathbb{R}^n$  is a point  $\text{proj}_S(\mathbf{y}) \in S$ :

$$\text{proj}_S(\mathbf{y}) = \frac{1}{2} \underset{\mathbf{x} \in S}{\text{argmin}} \|\mathbf{x} - \mathbf{y}\|_2^2$$

- **Sufficient conditions of existence of a projection.** If  $S \subseteq \mathbb{R}^n$  - closed set, then the projection on set  $S$  exists for any point.
- **Sufficient conditions of uniqueness of a projection.** If  $S \subseteq \mathbb{R}^n$  - closed convex set, then the projection on set  $S$  is unique for any point.
- If a set is open, and a point is beyond this set, then its projection on this set does not exist.
- If a point is in set, then its projection is the point itself.

# Projection

## 💡 Bourbaki-Cheney-Goldstein inequality theorem

Let  $S \subseteq \mathbb{R}^n$  be closed and convex,  $\forall x \in S, y \in \mathbb{R}^n$ . Then

$$\langle y - \text{proj}_S(y), x - \text{proj}_S(y) \rangle \leq 0 \quad (1)$$

$$\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2 \quad (2)$$

## 💡 Non-expansive function

A function  $f$  is called **non-expansive** if  $f$  is  $L$ -Lipschitz with  $L \leq 1$ . That is, for any two points  $x, y \in \text{dom} f$ ,

$$\|f(x) - f(y)\| \leq L\|x - y\|, \text{ where } L \leq 1.$$

It means the distance between the mapped points is possibly smaller than that of the unmapped points.

Non-expansive becomes contractive if  $L < 1$ .

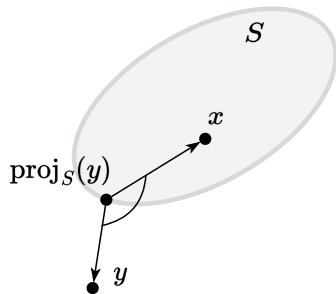


Figure 1: Obtuse or straight angle should be for any point  $x \in S$

# Problems

## Question

Is projection operator non-expansive?

## Question

Find projection  $\text{proj}_S(\mathbf{y})$  onto  $S$ , where  $S$ :

- $l_2$ -ball with center 0 and radius 1:

$$S = \{x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^d x_i^2 \leq 1\}$$

- $\mathbb{R}^d$ -cube:

$$S = \{x \in \mathbb{R}^d \mid a_i \leq x_i \leq b_i\}$$

- Affine constraints:

$$S = \{x \in \mathbb{R}^d \mid Ax = b\}$$

## Projected Gradient Descent (PGD). Idea

$$x_{k+1} = \text{proj}_S(x_k - \alpha_k \nabla f(x_k)) \quad \Leftrightarrow \quad \begin{aligned} y_k &= x_k - \alpha_k \nabla f(x_k) \\ x_{k+1} &= \text{proj}_S(y_k) \end{aligned}$$

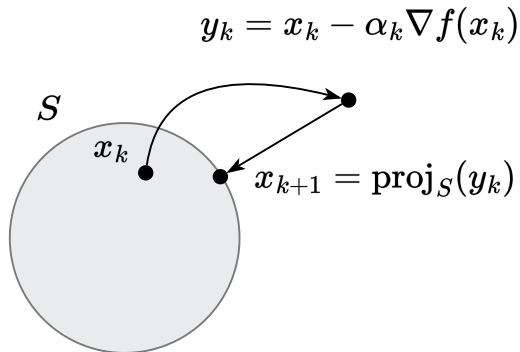


Figure 2: Illustration of Projected Gradient Descent algorithm

## Frank-Wolfe Method (FWM). Idea

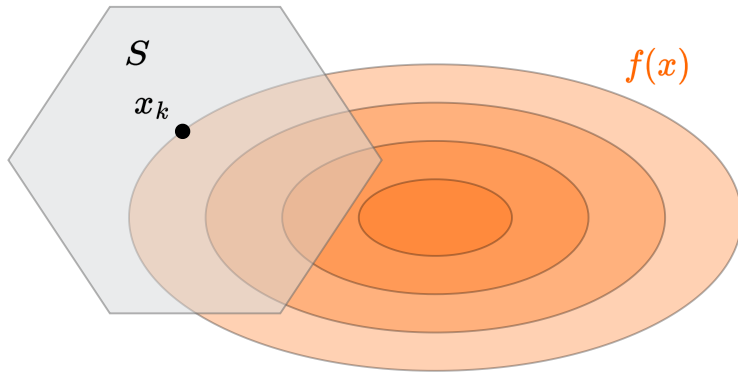


Figure 3: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

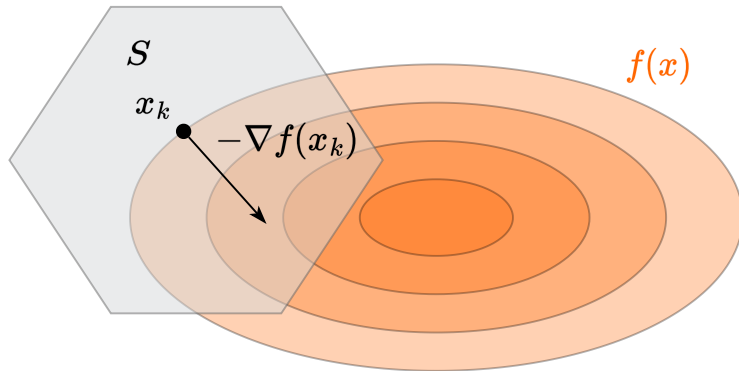


Figure 4: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

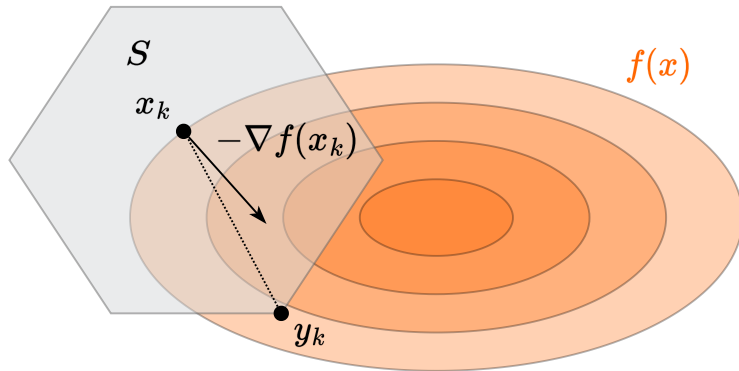


Figure 5: Illustration of Frank-Wolfe (conditional gradient) algorithm



## Frank-Wolfe Method (FWM). Idea

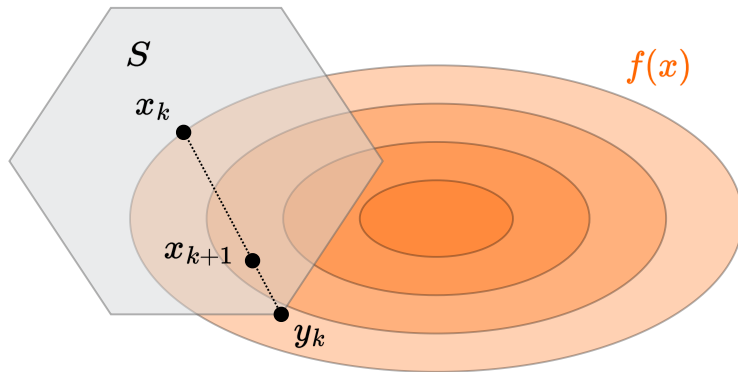


Figure 6: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

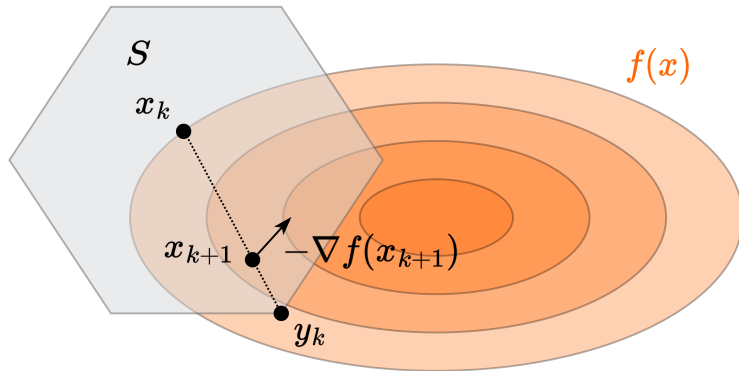


Figure 7: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

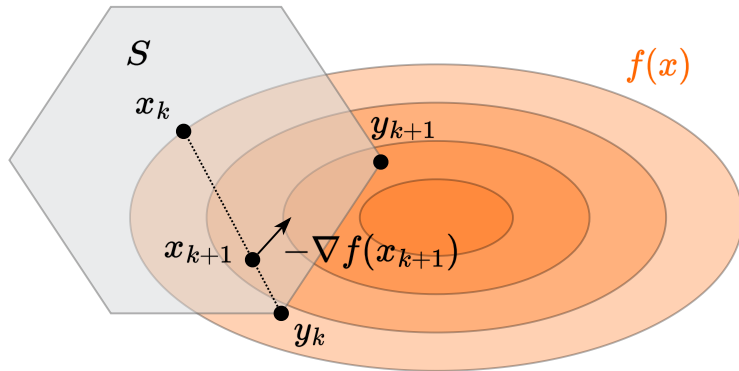


Figure 8: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

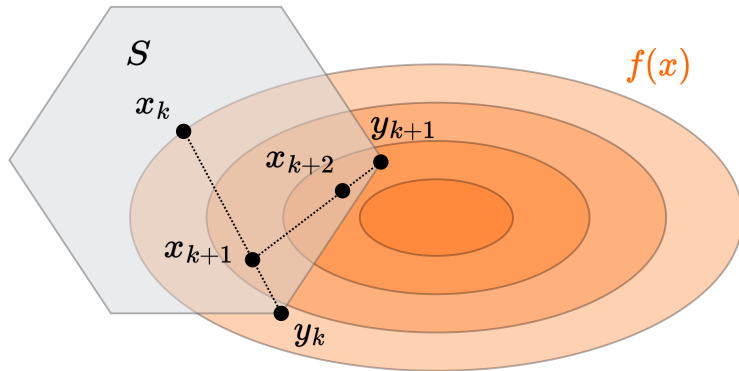


Figure 9: Illustration of Frank-Wolfe (conditional gradient) algorithm

## Frank-Wolfe Method (FWM). Idea

$$y_k = \arg \min_{x \in S} f_{x_k}^I(x) = \arg \min_{x \in S} \langle \nabla f(x_k), x \rangle$$

$$x_{k+1} = \gamma_k x_k + (1 - \gamma_k) y_k$$

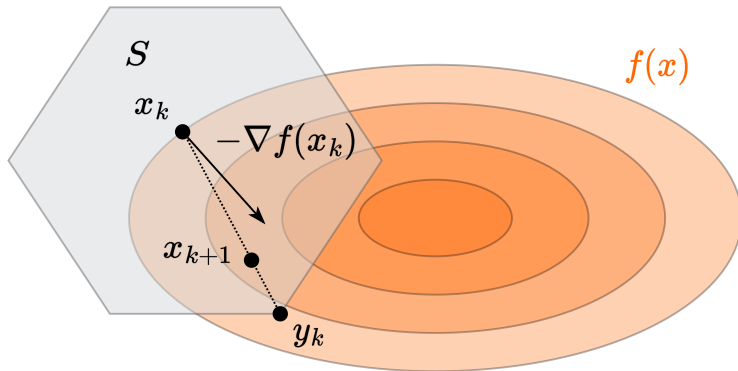


Figure 10: Illustration of Frank-Wolfe (conditional gradient) algorithm

# Convergence rate for smooth and convex case

## Theorem

Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and differentiable. Let  $S \subseteq \mathbb{R}^n$  be a closed convex set, and assume that there is a minimizer  $x^*$  of  $f$  over  $S$ ; furthermore, suppose that  $f$  is smooth over  $S$  with parameter  $L$ .

- The **Projected Gradient Descent** algorithm with stepsize  $\frac{1}{L}$  achieves the following convergence after iteration  $k > 0$ :

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}$$

- The **Frank-Wolfe Method** achieves the following convergence after iteration  $k > 0$ :

$$f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{k+1}$$

## 💡 FWM specificity

- FWM convergence rate for the  $\mu$ -strongly convex functions is  $\mathcal{O}\left(\frac{1}{k}\right)$
- FWM doesn't work for non-smooth functions. But modifications do.
- FWM works for any norm.

## Subgradient method: linear approximation + proximity

Recall SubGD step with sub-gradient  $g_k$ :

$$\begin{aligned} x_{k+1} &= \operatorname{argmin}_x \underbrace{f(x_k) + g_k^\top (x - x_k)}_{\text{linear approximation to } f} + \underbrace{\frac{1}{2\alpha} \|x - x_k\|_2^2}_{\text{proximity term}} \\ x_{k+1} &= x_k - \alpha_k g_k \quad \Leftrightarrow \\ &= \operatorname{argmin}_x \alpha g_k^\top x + \frac{1}{2} \|x - x_k\|_2^2 \end{aligned}$$

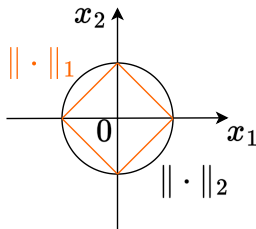


Figure 11:  $\|\cdot\|_1$  is not spherical symmetrical

## Example. Poor condition

Consider  $f(x_1, x_2) = x_1^2 \cdot \frac{1}{100} + x_2^2 \cdot 100$ .

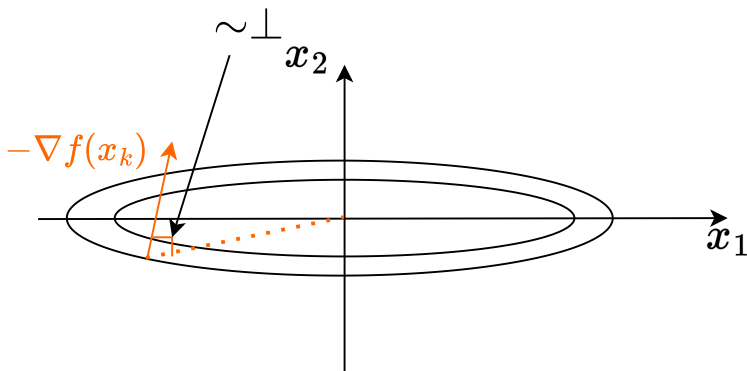


Figure 12: Poorly conditioned problem in  $\|\cdot\|_2$  norm



## Example. Poor condition

Suppose we are at the point:  $x_k = (-10 \quad -0.1)^\top$ . SubGD method:  $x_{k+1} = x_k - \alpha \nabla f(x_k)$

$$\nabla f(x_k) = \left( \frac{2x_1}{100} \quad 2x_2 \cdot 100 \right)^\top \Big|_{(-10 \quad -0.1)^\top} = \left( -\frac{1}{5} \quad -20 \right)^\top$$

**The problem:** due to elongation of the level sets the direction of movement  $(x_{k+1} - x_k)$  is  $\sim \perp (x^* - x_k)$ .

**The solution:** Change proximity term

$$x_{k+1} = \operatorname{argmin}_x \underbrace{f(x_k) + g_k^\top(x - x_k)}_{\text{linear approximation to } f} + \underbrace{\frac{1}{2\alpha}(x - x_k)^\top I(x - x_k)}_{\text{proximity term}}$$

to another

$$x_{k+1} = \operatorname{argmin}_x \underbrace{f(x_k) + g_k^\top(x - x_k)}_{\text{linear approximation to } f} + \underbrace{\frac{1}{2\alpha}(x - x_k)^\top Q(x - x_k)}_{\text{proximity term}},$$

where  $Q = \begin{pmatrix} \frac{1}{50} & 0 \\ 0 & 200 \end{pmatrix}$  for this example. And more generally to another function  $B_\phi(x, y)$  that measures proximity.

## Example. Poor condition

Let's find  $x_{k+1}$  for this **new** algorithm

$$\alpha \nabla f(x_k) + \begin{pmatrix} \frac{1}{50} & 0 \\ 0 & 200 \end{pmatrix} (x - x_k) = 0.$$

Solving for  $x$ , we get

$$x_{k+1} = x_k - \alpha \begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{200} \end{pmatrix} \nabla f(x_k) = (-10 \ -0.1)^\top - \alpha(-10 \ -0.1)^\top$$

**Observation:** Changing the proximity term, we **change the direction**  $x_{k+1} - x_k$ . In other words, if we measure distance using this **new** way, we also **change Lipschitzness**.

### Question

What is the Lipschitz constant of  $f$  at the point  $(1 \ 1)^\top$  for the norm:

$$\|z\|^2 = z^\top \begin{pmatrix} 50 & 0 \\ 0 & \frac{1}{200} \end{pmatrix} z?$$

## Example. Robust Regression

Square loss  $\|Ax - b\|_2^2$  is very sensitive to outliers.

**Instead:**  $\min \|Ax - b\|_1$ . This problem also **convex**.

Let's compute  $L$ -Lipshitz constant for  $f(x) = \|Ax - b\|_1$ :

$$|\|Ax - b\|_1 - \|Ay - b\|_1| \leq L\|x - y\|_2.$$

To simplify calculation:  $A = I$ ,  $b = 0$ , i.e.  $f(x) = \|x\|_1$ .

If we take  $x = \mathbf{1}_d$ ,  $y = (1 + \varepsilon)\mathbf{1}_d$ :


$$|n - (1 + \varepsilon)n| = \varepsilon n \leq L\|x - y\|_2 = \|\varepsilon\|_2 = \sqrt{n\varepsilon^2} = \varepsilon\sqrt{n}.$$

Finally, we get  $L = \sqrt{n}$ . As we can see,  $L$  is **dimension dependent**.

### Question

Show that if  $\|\nabla f(x)\|_\infty \leq 1$ , then  $\|\nabla f(x)\|_2 \leq \sqrt{d}$ .

# References

Examples for the Mirror Descent was taken from the  Lecture.