Optimization on "simple" sets. Gradient projection method. Conditional gradient method

Optimization in ML

Aleksandr Beznosikov

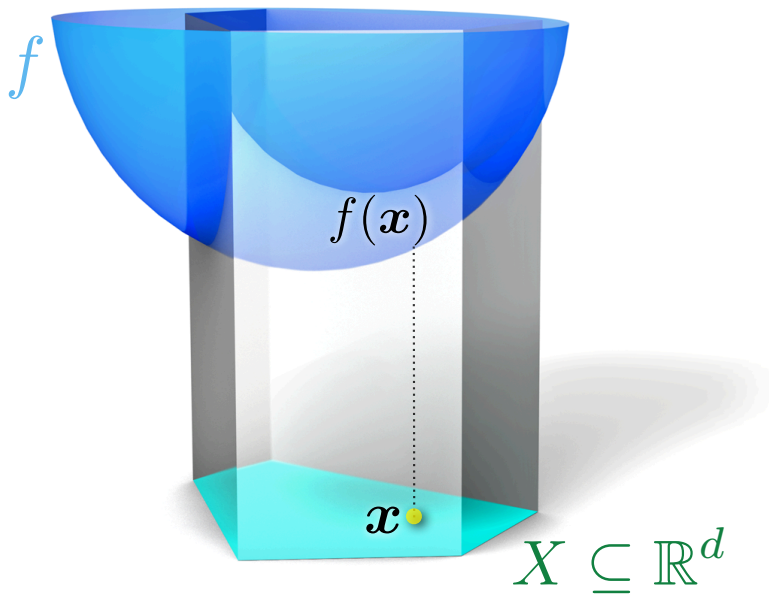
Skoltech

28 November 2023

Constrained Optimization

Constrained Optimization Problem

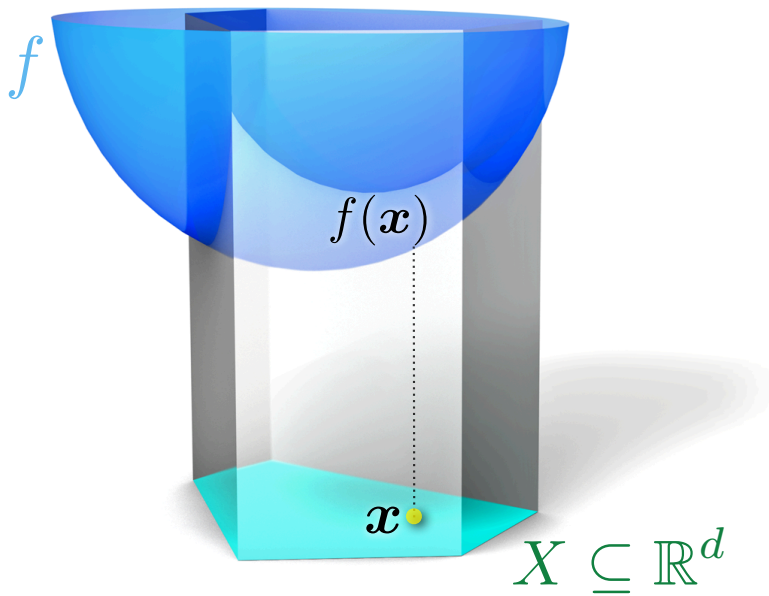
$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in X \end{aligned}$$



Constrained Optimization

Constrained Optimization Problem

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

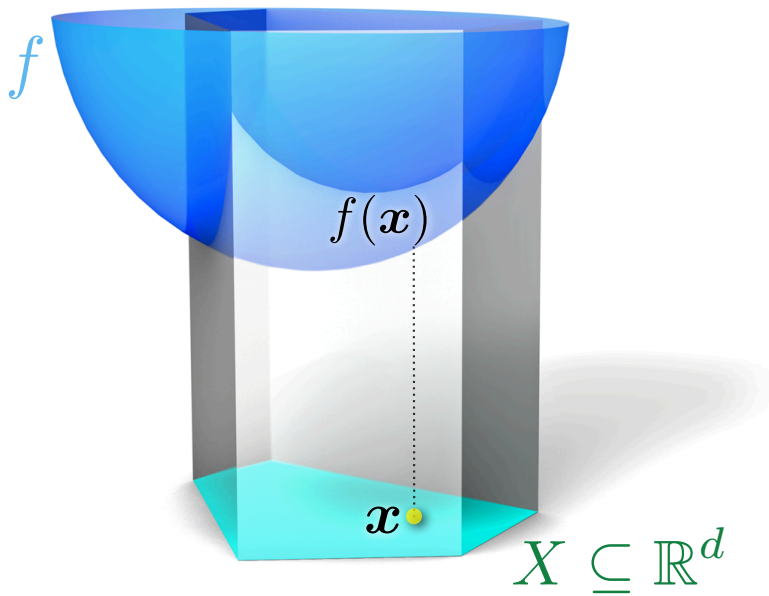


- Before $X = \mathbb{R}^d$ (unconstrained optimization)

Constrained Optimization

Constrained Optimization Problem

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

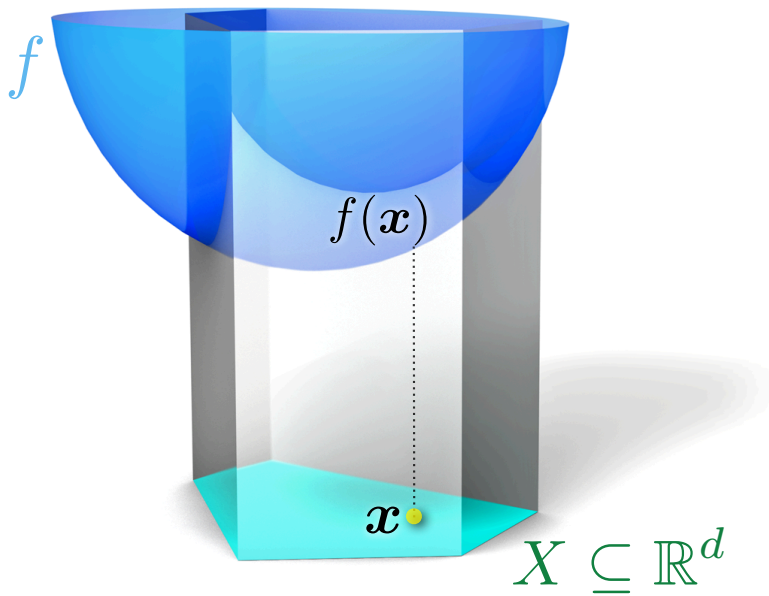


- Before $X = \mathbb{R}^d$ (unconstrained optimization)
- This lecture: $X \subsetneq \mathbb{R}^d$ (convex set)

Constrained Optimization

Constrained Optimization Problem

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



- Before $X = \mathbb{R}^d$ (unconstrained optimization)
- This lecture: $X \subsetneq \mathbb{R}^d$ (convex set)

Question: Why do we need constrains?

Example: PhD's Admission

- Skoltech university is admitting top students to its PhD program, in a competitive application process.
- Applicants are submitting various documents (GPA, TOEFL,...)
- Admission committee would like to compute a (rough) forecast of the applicant's performance in the program, based on the submitted documents.
- Data on the actual performance of students admitted in the past is available.
- In the following (made-up toy) example: consider GPA and TOEFL only...
- ...as predictors for Performance (final score of success obtained in the program).

Example: PhD's Admission

- $0.0 \leq$ GPA ≤ 4.0 (from F to A)
- $0 \leq$ TOEFL
- $1.0 \leq$ Performance ≤ 6.0 (final score of secess)
- Historical data:

GPA	TOEFL	Performance
3.52	100	3.92
3.66	109	4.34
3.76	113	4.80
3.74	100	4.67
3.93	100	5.52
3.88	115	5.44
3.77	115	5.04
3.66	107	4.73
3.87	106	5.03
3.84	107	5.06

PhD's Admission: Linear model

Hypothesis:

$$\text{Performance} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{TOEFL}$$

for weights w_0, w_1, w_2 to be learned.
Approach: Find w_0, w_1, w_2 by minimizing least squares error over the historical data.
Question: what we need to do with data before solving something?

PhD's Admission: Linear model

Hypothesis:

$$\text{Performance} \approx w_0 + w_1 \cdot \text{GPA} + w_2 \cdot \text{TOEFL}$$

for weights w_0, w_1, w_2 to be learned.
Approach: Find w_0, w_1, w_2 by minimizing least squares error over the historical data.
Question: what we need to do with data before solving something?

- Relevant GPA scores span a range of 0.5.
- Relevant TOEFL scores span a range of 20 (from 100 to 120).
- $\Rightarrow$ normalize first so that w_1, w_2 can be compared

General setting

n inputs $x_1, \ldots, x_n$, $x_i \in \mathbb{R}^d$ for all i
d input variables $1, 2, \ldots, d$

- 10 (GPA, TOEFL) pairs, two input variables

n outputs $y_1, \ldots, y_n \in \mathbb{R}$

- 10 Perfomance scores

(x_i, y_i): an observation

- $((3.93, 100), 5.52)$, observation (of a student doing very well)

With weights $w_0, \mathrm{w} = (w_1, \ldots, w_d) \in \mathbb{R}^d$, we plan to minimize the least squares objective

$$f(w_0, \mathrm{w}) = \sum_{i=1}^{n} (w_0 + \mathrm{w}^T x_i - y_i)^2.$$

General setting: centering

Want to assume that

$$\frac{1}{n}\sum_{i=1}^{n} x_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0.$$

Can be achieved by

- subtracting the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ from every input
- subtracting the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ from every output.

Question: after centering what we can assume?

General setting: centering

Want to assume that

$$\frac{1}{n}\sum_{i=1}^{n} x_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} y_i = 0.$$

Can be achieved by

- subtracting the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ from every input
- subtracting the mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ from every output.

Question: after centering what we can assume?

After centering: $w_0^* = 0, \mathrm{w}^*$ is unaffected

$\Rightarrow$ From now on consider function

$$f(\mathrm{w}) = \sum_{i=1}^{n} (\mathrm{w}^T x_i - y_i)^2.$$

General setting: normalization

Want to assume that for all j, the n input values $x_{1j}, \dots x_{nj}$ are on the same scale:

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1, \quad j = 1, \dots, d.$$

Can be achieved by

- multiplying x_{ij} by $s(j) = \sqrt{n/\sum_{i=1}^{n} x_{ij}^2}$ for all i, j
- in w*, this just multiplies w_j^* by $1/s(j)$

PhD's Admission: Centered and normalized data

x_{i1} (GPA)	x_{i2} (TOEFL)	y_i (Perfomance)
-2.04	-1.28	-0.94
-0.88	0.32	-0.52
-0.05	1.03	-0.05
-0.16	-1.28	-0.18
1.42	-1.28	0.67
1.02	1.39	0.59
0.06	1.39	0.19
-0.88	-0.04	-0.12
0.89	-0.21	0.17
0.62	-0.04	0.21

Least-squares objective:

$$f(w_1, w_2) = \sum_{i=1}^{10} (w_1 x_{i1} + w_2 x_{i2} - y_i)^2.$$

PhD's Admission: Results

Optimal solution: $\mathrm{w}^* = (w_1^*, w_2^*) \approx (0.43, 0.097)$

PhD's Admission: Results

Optimal solution: $\mathrm{w}^* = (w_1^*, w_2^*) \approx (0.43, 0.097)$

Under hypothesis (linear model), we expect $y_i \approx y_i^* = 0.43x_{i1} + 0.097x_{i2}$

x_{i1}	x_{i2}	y_i	y_i^*
-2.04	-1.28	-0.94	-1.00
-0.88	0.32	-0.52	-0.35
-0.05	1.03	-0.05	0.08
-0.16	-1.28	-0.18	-0.19
1.42	-1.28	0.67	0.49
1.02	1.39	0.59	0.57
0.06	1.39	0.19	0.16
-0.88	-0.04	-0.12	-0.38
0.62	-0.04	0.21	0.26

Question: what we can say about results? TOEFL has only very small influence ($w_2^* = 0.097$)

Predicting Performance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Predicting Performance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution

Predicting Performance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Question:** easy of not?

Predicting Performance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Question:** easy of not? Non-convex or NP-hard – various k might have to be tried.

Question: if we have 100 features, how many different subsets (of features) can we have?

Predicting Performance in the future

Problems:

- least squares solution is optimized for the training data, not for the future ("overfitting")
- "unimportant" variables should have weight 0, but they typically don't

Subset selection heuristics: drop variables with seemingly "small" contribution (various methods to decide what "small" means, and how many to drop)

Best subset selection: solve least squares subject to an additional constraint that there are at most k nonzero weights. **Question:** easy of not? Non-convex or NP-hard – various k might have to be tried.

Question: if we have 100 features, how many different subsets (of features) can we have? $2^{100} \approx 1.26 \cdot 10^{30}$.

LASSO: popular approach with some favorable statistical properties

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$:

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: TOEFL is gone!

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: TOEFL is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$

The LASSO: a constrained optimization problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

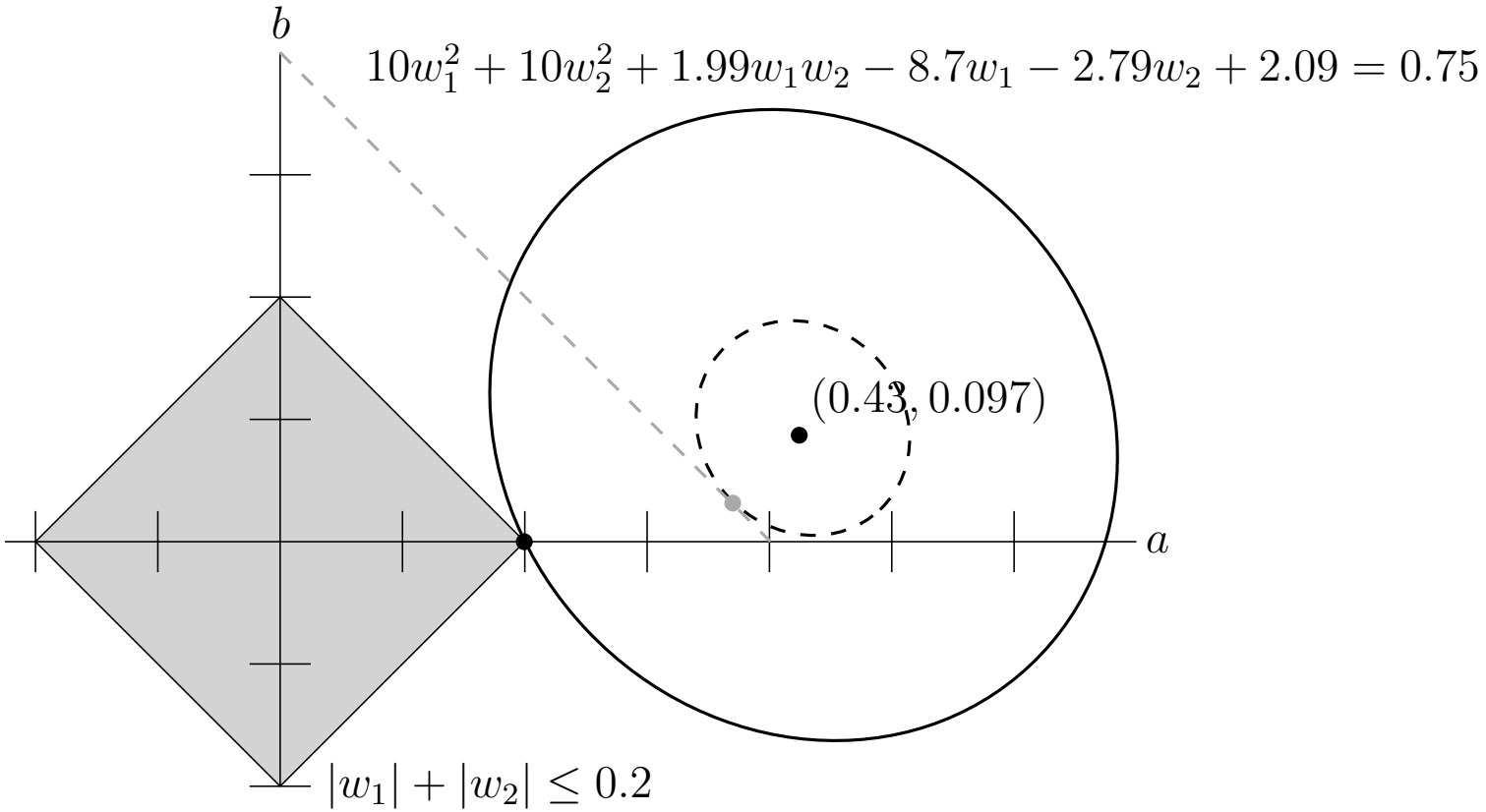
where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: TOEFL is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$
$R = 0.4 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.36, 0.036)$

The LASSO: a constrained optimization problem

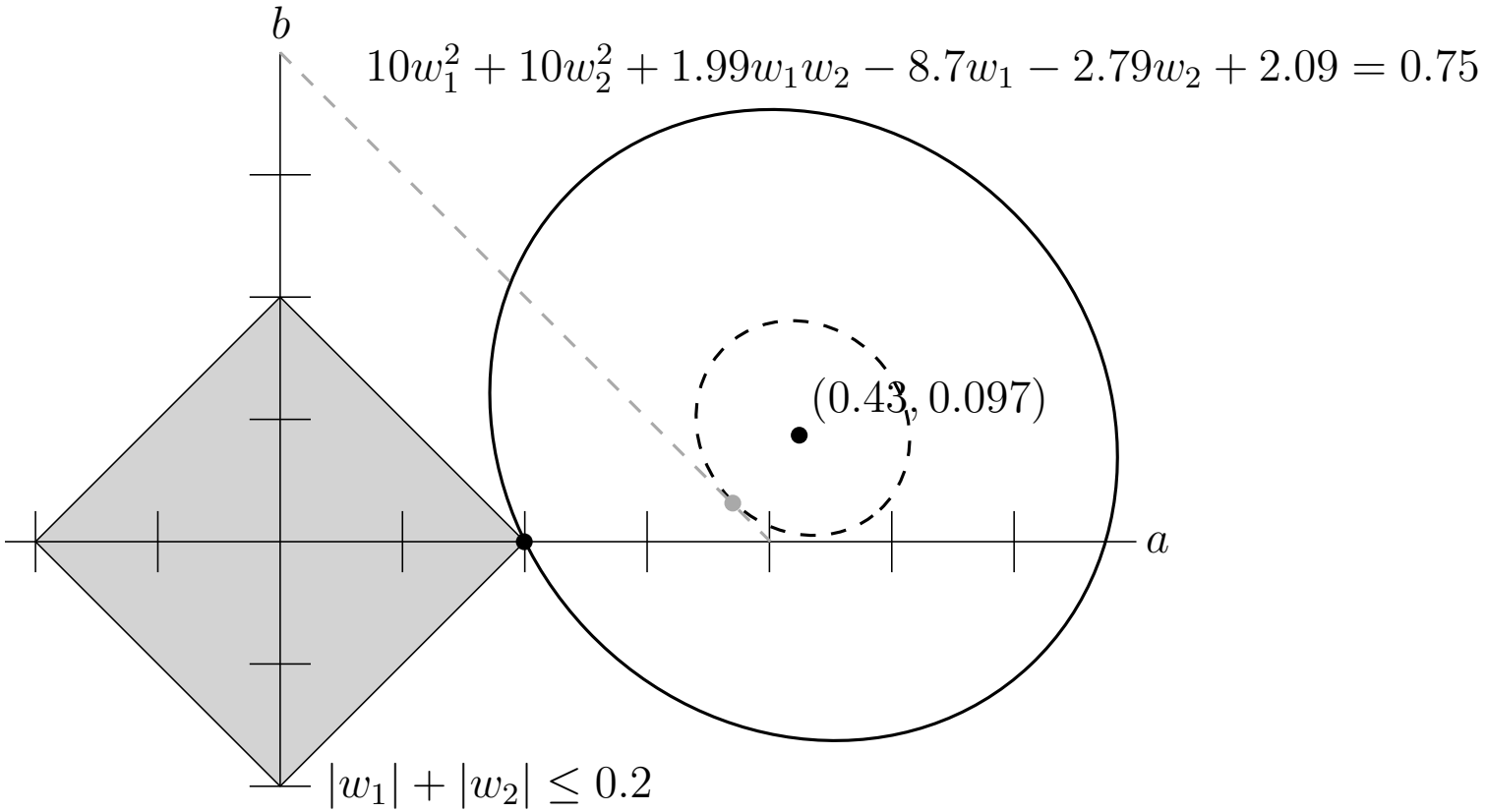
$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n} \|\mathrm{w}^\top x_i - y_i\|^2 \\ \text{subject to} & \|\mathrm{w}\|_1 \leq R, \end{array} \tag{1}$$

where $R \in \mathbb{R}_+$ is some parameter.
$\|\mathrm{w}\|_1 = \sum_{i=1}^{d} |w_j|$ is the 1-norm.
In our case:
$R = 0.2 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.2, 0)$: TOEFL is gone!
$R = 0.3 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.3, 0)$
$R = 0.4 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.36, 0.036)$
$R \geq 0.6 \Rightarrow \mathrm{w}^* = (w_1^*, w_2^*) = (0.43, 0.097)$

Geometry of the LASSO



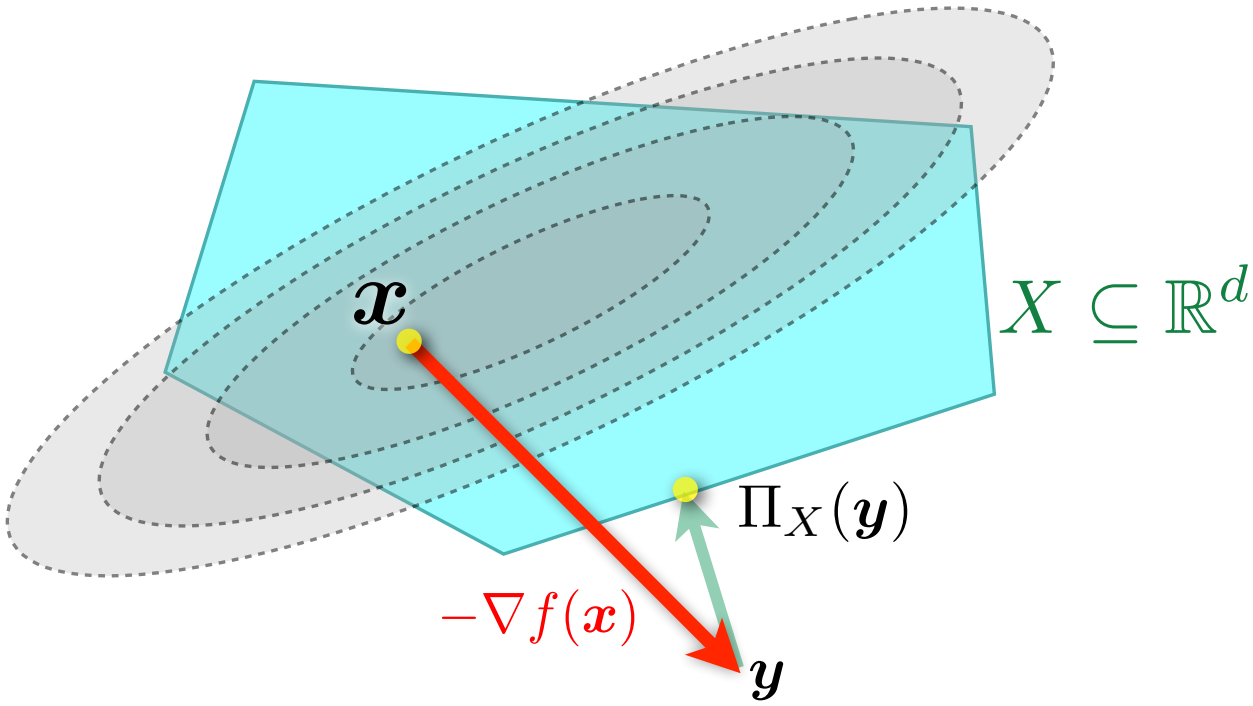
Geometry of the LASSO



Question: Can we somehow modify gradient method to work with constraints?

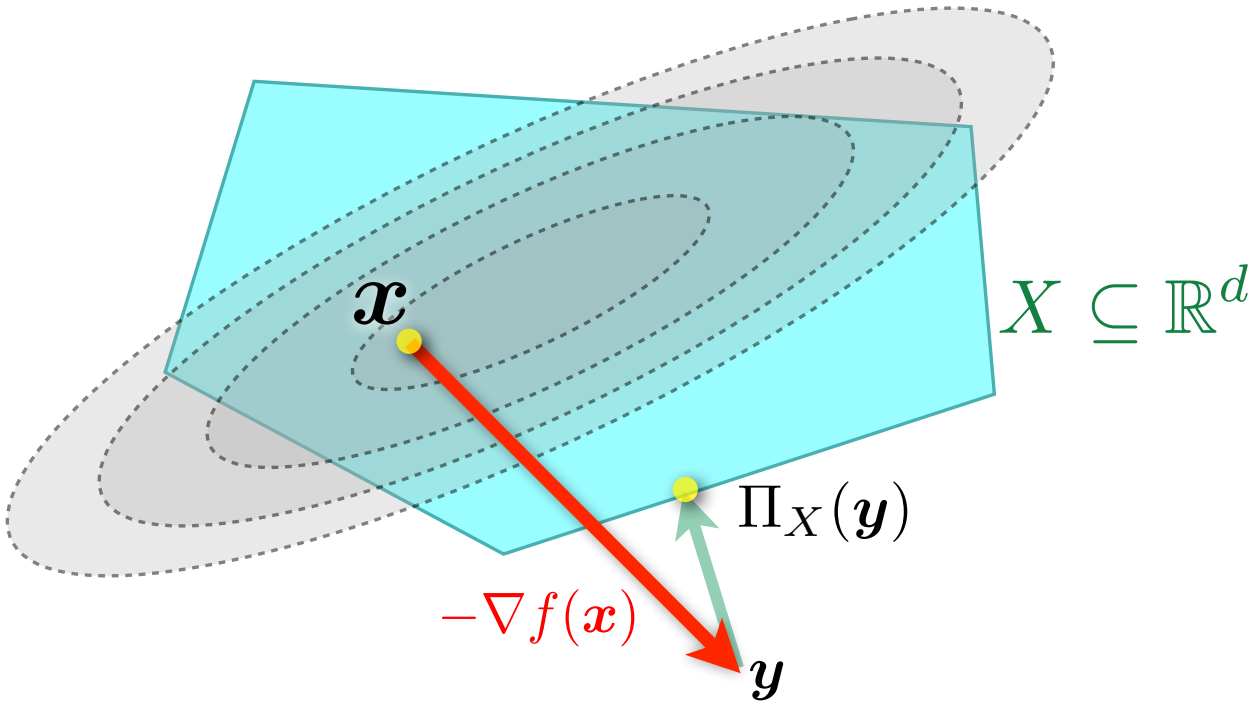
Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(y) := \operatorname{argmin}_{x \in X} \|x - y\|^2$



Projected Gradient Descent

Idea: project onto X after every step: $\Pi_X(y) := \operatorname{argmin}_{x \in X} \|x - y\|^2$



Projected gradient descent: $x_{k+1} := \Pi_X\left[x_k - \gamma \nabla f(x_k)\right]$

The Algorithm

Projected gradient descent:

The Algorithm

Projected gradient descent: choose $x_0 \in \mathbb{R}^d$.

The Algorithm

Projected gradient descent: choose $x_0 \in \mathbb{R}^d$.

$$y_{k+1} := x_k - \gamma \nabla f(x_k),$$

The Algorithm

Projected gradient descent: choose $x_0 \in \mathbb{R}^d$.

$$\begin{aligned} y_{k+1} &:= x_k - \gamma \nabla f(x_k), \\ x_{k+1} &:= \Pi_X(y_{k+1}) := \operatorname*{argmin}_{x \in X} \|x - y_{k+1}\|^2 \end{aligned}$$

The Algorithm

Projected gradient descent: choose $x_0 \in \mathbb{R}^d$.

$$\begin{aligned} y_{k+1} &:= x_k - \gamma \nabla f(x_k), \\ x_{k+1} &:= \Pi_X(y_{k+1}) := \operatorname*{argmin}_{x \in X} \|x - y_{k+1}\|^2 \end{aligned}$$

for times $k = 0, 1, \ldots,$ and stepsize $\gamma \geq 0$.

Optimality condition for the problem with constraints

Let there be a convex continuously differentiable on $\mathbb{R}^d$ function $f : \mathbb{R}^d \to \mathbb{R}$ and the convex set $\mathcal{X}$. Then $x^* \in \mathcal{X}$ is a global minimum of f on $\mathcal{X}$ if and only if for all $x \in \mathcal{X}$ the following holds

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0.$$

Proof

- Sufficient condition.

Proof

- Sufficient condition. Let $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for $x \in \mathcal{X}$, then we use the definition of convexity:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

Whence it follows that x^* — the global minimum on $\mathcal{X}$.

Proof

- Sufficient condition. Let $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for $x \in \mathcal{X}$, then we use the definition of convexity:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Whence it follows that x^* — the global minimum on $\mathcal{X}$.
- Necessary condition.

Proof

- Sufficient condition. Let $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for $x \in \mathcal{X}$, then we use the definition of convexity:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Whence it follows that x^* — the global minimum on $\mathcal{X}$.
- Necessary condition. Let x^* be a global minimum on $\mathcal{X}$. We prove that $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for any $x \in \mathcal{X}$. Let us proceed from the contrary, i.e. suppose that there exists $x \in \mathcal{X}$ such that $\langle \nabla f(x^*), x - x^* \rangle < 0$.

Proof

- Sufficient condition. Let $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for $x \in \mathcal{X}$, then we use the definition of convexity:

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^* \rangle \geq f(x^*).$$

  Whence it follows that x^* — the global minimum on $\mathcal{X}$.

- Necessary condition. Let x^* be a global minimum on $\mathcal{X}$. We prove that $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for any $x \in \mathcal{X}$. Let us proceed from the contrary, i.e. suppose that there exists $x \in \mathcal{X}$ such that $\langle \nabla f(x^*), x - x^* \rangle < 0$. Let us consider

$$x_\lambda = \lambda x + (1 - \lambda)x^*, \text{ where } \lambda \in [0; 1].$$

  Due to the convexity of the set $\mathcal{X}$, the points $x_\lambda \in \mathcal{X}$.

Proof

Let us understand how the function $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$ behaves. In particular, note that

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Proof

Let us understand how the function $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$ behaves. In particular, note that

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Also note that $\frac{d\phi}{d\lambda}\big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$.

Proof

Let us understand how the function $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$ behaves. In particular, note that

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Also note that $\frac{d\phi}{d\lambda}\big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Question:** what does this mean?

Proof

Let us understand how the function $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$ behaves. In particular, note that

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Also note that $\frac{d\phi}{d\lambda}\big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Question:** what does this mean? the function ϕ is decreasing in the neighborhood of zero. It means that for small enough $\lambda > 0$, the following is true

$$f(x^* + \lambda(x - x^*)) = \phi(\lambda) < \phi(0) = f(x^*).$$

Proof

Let us understand how the function $\phi(\lambda) = f(x_\lambda) = f(\lambda x + (1-\lambda)x^*)$ behaves. In particular, note that

$$\frac{d\phi}{d\lambda} = \frac{d}{d\lambda} f(\lambda x + (1-\lambda)x^*) = \langle \nabla f(x^* + \lambda(x - x^*)), x - x^* \rangle.$$

Also note that $\frac{d\phi}{d\lambda}\big|_{\lambda=0} = \langle \nabla f(x^*), x - x^* \rangle < 0$. **Question:** what does this mean? the function ϕ is decreasing in the neighborhood of zero. It means that for small enough $\lambda > 0$, the following is true

$$f(x^* + \lambda(x - x^*)) = \phi(\lambda) < \phi(0) = f(x^*).$$

Come to the contradiction that x^* — a global minimum on $\mathcal{X}$.

Properties of the projection operator

Property of the projection operator

For a convex set $\mathcal{X}$ and any point, the projection operator exists and takes a singular value.

Properties of the projection operator

Property of the projection operator

For a convex set $\mathcal{X}$ and any point, the projection operator exists and takes a singular value.

Proof:

Properties of the projection operator

Property of the projection operator

For a convex set $\mathcal{X}$ and any point, the projection operator exists and takes a singular value.

Proof: It follows that the projection problem is the minimization of a strongly convex function on a convex set. The solution of such a problem exists and is unique.

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Proof:

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Proof: Note that $\Pi_{\mathcal{X}}(y)$ minimizes the differentiable convex function $d(z) = \|z - y\|_2^2$ on the convex set $\mathcal{X}$. **Question:** what then does the optimality condition give us?

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Proof: Note that $\Pi_{\mathcal{X}}(y)$ minimizes the differentiable convex function $d(z) = \|z - y\|_2^2$ on the convex set $\mathcal{X}$. **Question:** what then does the optimality condition give us?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \quad \text{для любой } x \in \mathcal{X}$$

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Proof: Note that $\Pi_{\mathcal{X}}(y)$ minimizes the differentiable convex function $d(z) = \|z - y\|_2^2$ on the convex set $\mathcal{X}$. **Question:** what then does the optimality condition give us?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \quad \text{для любой } x \in \mathcal{X}$$

Question: what is $\nabla d(z)$?

Properties of the projection operator

Property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x \in \mathcal{X}, y \in \mathbb{R}^d$. Then

$$\langle x - \Pi_{\mathcal{X}}(y), y - \Pi_{\mathcal{X}}(y) \rangle \leq 0.$$

Proof: Note that $\Pi_{\mathcal{X}}(y)$ minimizes the differentiable convex function $d(z) = \|z - y\|_2^2$ on the convex set $\mathcal{X}$. **Question:** what then does the optimality condition give us?

$$\langle \nabla d(\Pi_{\mathcal{X}}(y)), x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \quad \text{для любой } x \in \mathcal{X}$$

Question: what is $\nabla d(z)$? $2(z - y)$. Then

$$2\langle \Pi_{\mathcal{X}}(y) - y, x - \Pi_{\mathcal{X}}(y) \rangle \geq 0 \quad \text{for all } x \in \mathcal{X}.$$

Which is exactly what I needed to prove.

Properties of the projection operator

Non-expansivity property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x_1, x_2 \in \mathbb{R}^d$. Then

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Properties of the projection operator

Non-expansivity property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x_1, x_2 \in \mathbb{R}^d$. Then

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Proof:

Properties of the projection operator

Non-expansivity property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x_1, x_2 \in \mathbb{R}^d$. Then

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Proof: From the previous property for $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

Properties of the projection operator

Non-expansivity property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x_1, x_2 \in \mathbb{R}^d$. Then

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Proof: From the previous property for $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

Substitute $x = \Pi_{\mathcal{X}}(x_2)$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

Properties of the projection operator

Non-expansivity property of the projection operator

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set, $x_1, x_2 \in \mathbb{R}^d$. Then

$$\|\Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\|_2 \leq \|x_1 - x_2\|_2.$$

Proof: From the previous property for $x \in \mathcal{X}$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), x - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

Substitute $x = \Pi_{\mathcal{X}}(x_2)$:

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\rangle \leq 0$$

The same:

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2)\rangle \leq 0$$

Properties of the projection operator

Sum it up

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

and

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

Properties of the projection operator

Sum it up

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

and

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

And we get

$$\langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq \langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), x_2 - x_1 \rangle$$

Properties of the projection operator

Sum it up

$$\langle x_1 - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq 0$$

and

$$\langle x_2 - \Pi_{\mathcal{X}}(x_2), \Pi_{\mathcal{X}}(x_1) - \Pi_{\mathcal{X}}(x_2) \rangle \leq 0$$

And we get

$$\langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1) \rangle \leq \langle \Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1), x_2 - x_1 \rangle$$

CBS gives the result we want

$$\|\Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\|_2^2 \leq \|\Pi_{\mathcal{X}}(x_2) - \Pi_{\mathcal{X}}(x_1)\| \cdot \|x_2 - x_1\|$$

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Proof:

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Proof: Let us rewrite:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right. \\ &\qquad \left. + \gamma^2 \|\nabla f(x^*)\|_2^2\right]\end{aligned}$$

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Proof: Let us rewrite:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right. \\ &\qquad \left. + \gamma^2 \|\nabla f(x^*)\|_2^2\right]\end{aligned}$$

From where

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle\right]$$

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Proof: Let us rewrite:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right. \\ &\qquad \left. + \gamma^2 \|\nabla f(x^*)\|_2^2\right]\end{aligned}$$

From where

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle\right]$$

Properties of the projection operator

Property of the projection operator

For x^* – the solution of the conditional problem of minimization of a convex continuously differentiable function f on the convex set $\mathcal{X}$ it is true that

$$x^* = \Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*))$$

Proof: Let us rewrite:

$$\begin{aligned}\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) &= \arg\min_{x \in \mathcal{X}} \|x^* - \gamma \nabla f(x^*) - x\|_2^2 \\ &= \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle \right. \\ &\qquad \left. + \gamma^2 \|\nabla f(x^*)\|_2^2\right]\end{aligned}$$

From where

$$\Pi_{\mathcal{X}}(x^* - \gamma \nabla f(x^*)) = \arg\min_{x \in \mathcal{X}} \left[\|x^* - x\|_2^2 + 2\gamma \langle \nabla f(x^*), x - x^* \rangle\right]$$

Proofs of convergence

- We're still considering it:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

Proofs of convergence

- We're still considering it:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

- Let us use the last proved property about the stationary point:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2 \\ &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - \Pi_{\mathcal{X}}\left[x^* - \gamma_k \nabla f(x^*)\right]\|_2^2\end{aligned}$$

Proofs of convergence

- We're still considering it:

$$\|x^{k+1} - x^*\|_2^2 = \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2$$

- Let us use the last proved property about the stationary point:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - x^*\|_2^2 \\ &= \|\Pi_{\mathcal{X}}\left[x^k - \gamma_k \nabla f(x^k)\right] - \Pi_{\mathcal{X}}\left[x^* - \gamma_k \nabla f(x^*)\right]\|_2^2\end{aligned}$$

- Now the third property

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq& \|x^k - \gamma_k \nabla f(x^k) - x^* + \gamma_k \nabla f(x^*)\|_2^2 \\ =& \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle \\ &+ \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Proof of convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

Proof of convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

- Let us introduce the following object the Bragman divergence generated by the convex function f:

$$V_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.$$

Question: why is divergence always positive?

Proof of convergence

- From the previous slide:

$$\|x^{k+1} - x^*\|_2^2 \leq \|x^k - x^*\|_2^2 - 2\gamma_k \langle \nabla f(x^k) - \nabla f(x^*), x^k - x^* \rangle + \gamma_k^2 \|\nabla f(x^k) - \nabla f(x^*)\|_2^2$$

- Let us introduce the following object the Bragman divergence generated by the convex function f:

$$V_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y \rangle \geq 0.$$

Question: why is divergence always positive? Because of the convexity of f.

Proof of convergence

- Let us take advantage of the strong convexity and smoothness as before:

$$\begin{aligned}\|x^{k+1} - x^*\|_2^2 \leq & \|x^k - x^*\|_2^2 \\ & - 2\gamma_k \left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle + \frac{\mu}{2} \|x^k - x^*\|_2^2 \right) \\ & + 2\gamma_k^2 L \left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle \right) \\ = & (1 - \mu\gamma_k) \|x^k - x^*\|_2^2 + 2\gamma_k(\gamma_k L - 1) V_f(x^k, x^*)\end{aligned}$$

- Further as before, due to the non-negativity of the Bragman divergence.

Convergence

- The gradient descent method with projection for a L-smooth and (μ-strongly) convex target function has the same convergence as the gradient descent method for a similar unconditional optimization problem.
- It turns out that the iterative and oracle complexities of these methods coincide.

Convergence

- The gradient descent method with projection for a L-smooth and (μ-strongly) convex target function has the same convergence as the gradient descent method for a similar unconditional optimization problem.
- It turns out that the iterative and oracle complexities of these methods coincide.
- One problem remains — projection is an additional optimization problem that needs to be solved.

Projecting is additional problem

- The projection has analytical solution for ℓ_2-ball with radius $R > 0$:

$$X = B_2(R) = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \leq R \right\}$$

$$\Pi_X(x) = \max\left\{1, \frac{R}{\|x\|_2}\right\} x$$

- For other quite simple sets we need to run additional optimization algorithms. For example, for $B_1(R)$ $\left(\left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}\right)$ the projection can be computed in time $\mathcal{O}(d \log d)$.

Projecting is additional problem

- The projection has analytical solution for ℓ_2-ball with radius $R > 0$:

$$X = B_2(R) = \left\{ x \in \mathbb{R}^d \mid \|x\|_2^2 = \sum_{i=1}^{d} x_i^2 \leq R \right\}$$

$$\Pi_X(x) = \max\left\{1, \frac{R}{\|x\|_2}\right\} x$$

- For other quite simple sets we need to run additional optimization algorithms. For example, for $B_1(R)$ $\left(\left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}\right)$ the projection can be computed in time $\mathcal{O}(d \log d)$.
- The next approach can help.

Linear minimization

Quadratic problem (projection) is hard... **Question:** what we can try instead?

Linear minimization

Quadratic problem (projection) is hard... **Question:** what we can try instead? Linear problems:

$$\min_{s \in X} \langle s, g \rangle$$

Question: Is it easy to solve?

- ℓ_1-ball: $B_1(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\}$

Linear minimization

Quadratic problem (projection) is hard... **Question:** what we can try instead? Linear problems:

$$\min_{s \in X} \langle s, g \rangle$$

Question: Is it easy to solve?

- ℓ_1-ball: $B_1(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\}$
- probability simplex: $\triangle = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \ \sum_{i=1}^{d} x_i = 1 \right\}$

Linear minimization

Quadratic problem (projection) is hard... **Question:** what we can try instead? Linear problems:

$$\min_{s \in X} \langle s, g \rangle$$

Question: Is it easy to solve?

- ℓ_1-ball: $B_1(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\}$
- probability simplex: $\triangle = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \sum_{i=1}^{d} x_i = 1 \right\}$
- ℓ_∞-ball: $B_\infty(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_\infty = \max_{i=1,\ldots d} |x_i| \leq 1 \right\}$

Linear minimization

Quadratic problem (projection) is hard... **Question:** what we can try instead? Linear problems:

$$\min_{s \in X} \langle s, g \rangle$$

Question: Is it easy to solve?

- ℓ_1-ball: $B_1(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_1 = \sum_{i=1}^{d} |x_i| \leq 1 \right\}$
- probability simplex: $\triangle = \left\{ x \in \mathbb{R}^d \mid x_i \geq 0, \ \sum_{i=1}^{d} x_i = 1 \right\}$
- ℓ_∞-ball: $B_\infty(1) = \left\{ x \in \mathbb{R}^d \mid \|x\|_\infty = \max_{i=1,\ldots d} |x_i| \leq 1 \right\}$

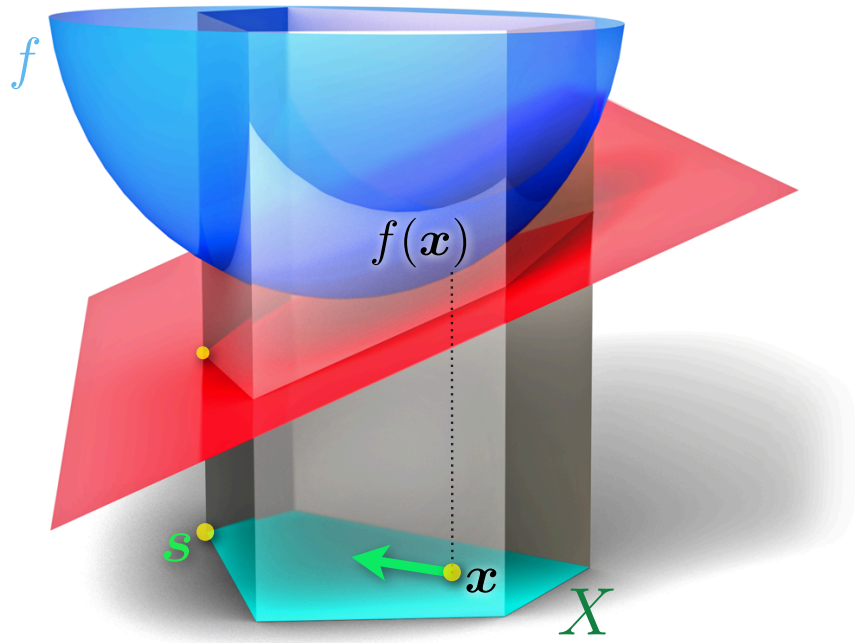
Examples	Solution of LM
ℓ_1-ball	$-\text{sign}(g_i)e_i$ with $\text{argmax}_i \lvert g_i \rvert$
Simplex	e_i with $\text{argmin}_i g_i$
ℓ_∞-ball	$-\sum_{i=1}^{d} \text{sign}(g_i)e_i$

Frank-Wolfe Algorithm

Frank-Wolfe Algorithm:

$$\begin{aligned} s &:= \text{LMO}(\nabla f(x_k)), \\ x_{k+1} &:= (1-\gamma)x_k + \gamma s, \end{aligned}$$

for timesteps $k = 0, 1, \ldots,$ and stepsize $\gamma := \frac{2}{k+2}$.



Linear Minimization Oracle:

$$\text{LMO}(g) := \operatorname*{argmin}_{s \in X} \langle s, g \rangle$$

Properties

- Aways feasible: $x_0, x_1, \ldots, x_k \in X$.
  x_{k+1} is on line segment $[s, x_k]$, for $\gamma \in [0, 1]$.
- Reduces non-linear to linear optimization
- Projection-free
- Sparse iterates (in terms of corners s used)

Invented and analyzed 1956 by Marguerite Frank and Philip Wolfe.
Check **Marguerite Frank - Honorary Discussion Panel @ NeurIPS 2013 workshop**
`https://www.youtube.com/watch?v=24e08AX9Eww`

Convergence in $\mathcal{O}(1/\varepsilon)$ steps

Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and smooth with parameter L, and $x_0 \in X$. Then choosing any of the above stepsizes, the Frank-Wolfe algorithm yields

$$f(x_K) - f(x^*) \leq \frac{\max\{2L\,\mathrm{diam}(X)^2, f(x_0) - f(x^*)\}}{K+1}$$

Where $\mathrm{diam}(X) := \max_{x,y \in X} \|x - y\|$ is the diameter of X.

Proof of Convergence in $\mathcal{O}(1/\varepsilon)$ steps

- L -smoothness gives

$$\begin{aligned} f(x_{k+1}) =& f(x_k + \gamma_k(s - x_k)) \\ \leq& f(x_k) + \gamma_k \langle s - x_k, \nabla f(x_k) \rangle + \frac{\gamma_k^2}{2} L \|s - x_k\|^2 \\ \leq& f(x_k) + \gamma_k \langle s - x_k, \nabla f(x_k) \rangle + \frac{\gamma_k^2}{2} L \operatorname{diam}(X)^2. \end{aligned}$$

- By LMO property and convexity gives

$$\begin{aligned} f(x_{k+1}) - f(x^*) \leq& f(x_k) - f(x^*) + \gamma_k \langle x^* - x_k, \nabla f(x_k) \rangle + \frac{\gamma_k^2}{2} L \operatorname{diam}(X)^2 \\ =& f(x_k) - f(x^*) - \gamma_k (f(x_k) - f(x^*)) + \frac{\gamma_k^2}{2} L \operatorname{diam}(X)^2 \\ =& (1 - \gamma_k)(f(x_k) - f(x^*)) + \gamma_k^2 C. \end{aligned}$$

Here $C = \frac{L \operatorname{diam}(X)^2}{2}$.

The induction step

We will now use induction over k to prove our claimed bound, i.e.

$$f(x_k) - f(x^*) \leq \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2} \quad k = 0, 1, \ldots$$

The induction step

We will now use induction over k to prove our claimed bound, i.e.

$$f(x_k) - f(x^*) \leq \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2} \quad k = 0, 1, \ldots$$

The base-case $k = 0$ follows automatically.
Now considering $k \geq 1$,

$$\begin{aligned} f(x_{k+1}) - f(x^*) \leq & \quad (1 - \gamma_k)(f(x_k) - f(x^*)) + {\gamma_t}^2 C \\ = & \quad \left(1 - \frac{2}{k+2}\right)(f(x_k) - f(x^*)) + \left(\frac{2}{k+2}\right)^2 C \\ \leq & \quad \left(1 - \frac{2}{k+2}\right) \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2} + \left(\frac{2}{k+2}\right)^2 C \ , \end{aligned}$$

where in the last inequality we have used the induction hypothesis of induction.

The induction step

Simply rearranging the terms gives

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2}\left(1 - \tfrac{2}{k+2} + \tfrac{1}{k+2}\right) \\ &= \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2}\left(1 - \tfrac{1}{k+2}\right) \\ &= \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2}\,\frac{k+2-1}{k+2} \\ &\leq \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+2}\,\frac{k+2}{k+3} \\ &= \frac{\max\{4C; f(x_0) - f(x^*)\}}{k+3}, \end{aligned}$$

which is our claimed bound for $k \geq 1$.

Convergence

Convergence theorem of the Frank-Wolfe method

Let a continuously differentiable convex L-smooth function $f : \mathbb{R}^d \to \mathbb{R}$ be given, then the following convergence estimate is valid for the Frank-Wolfe method

$$f(x^K) - f(x^*) \leq \frac{\max\{2L \operatorname{diam}(\mathcal{X})^2, f(x^0) - f(x^*)\}}{K + 2}$$

where $\operatorname{diam}(\mathcal{X}) := \max_{x,y \in \mathcal{X}} \|x - y\|$ – set $\mathcal{X}$ diameter.

Convergence

- The $1/K$ sublinear convergence is the same as that of gradient descent for a convex L-smooth function.

Convergence

- The $1/K$ sublinear convergence is the same as that of gradient descent for a convex L-smooth function.
- The problem that for a strongly convex target function, linear convergence will not appear in the general case. It is related to linear minimization.