



Gradient methods for conditional problems.
Projected Gradient Descent. Frank-Wolfe
method. Idea of Mirror Descent algorithm

Даня Меркулов

Методы Оптимизации в Машинном Обучении. ФКН ВШЭ

Conditional methods

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.
- The solution has to be inside the set S .

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.
- The solution has to be inside the set S .
- Example:

$$\frac{1}{2} \|Ax - b\|_2^2 \rightarrow \min_{\|x\|_2^2 \leq 1}$$

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.
- The solution has to be inside the set S .
- Example:

$$\frac{1}{2} \|Ax - b\|_2^2 \rightarrow \min_{\|x\|_2^2 \leq 1}$$

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.
- The solution has to be inside the set S .
- Example:

$$\frac{1}{2} \|Ax - b\|_2^2 \rightarrow \min_{\|x\|_2^2 \leq 1}$$

Gradient Descent is a great way to solve unconstrained problem

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad (\text{GD})$$

Is it possible to tune GD to fit constrained problem?

Constrained optimization

Unconstrained optimization

$$\min_{x \in \mathbb{R}^n} f(x)$$

- Any point $x_0 \in \mathbb{R}^n$ is feasible and could be a solution.

Constrained optimization

$$\min_{x \in S} f(x)$$

- Not all $x \in \mathbb{R}^n$ are feasible and could be a solution.
- The solution has to be inside the set S .
- Example:

$$\frac{1}{2} \|Ax - b\|_2^2 \rightarrow \min_{\|x\|_2^2 \leq 1}$$

Gradient Descent is a great way to solve unconstrained problem

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad (\text{GD})$$

Is it possible to tune GD to fit constrained problem?

Yes. We need to use projections to ensure feasibility on every iteration.

Example: White-box Adversarial Attacks

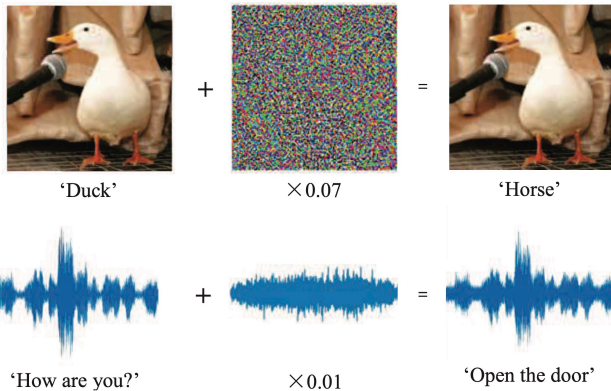


Рисунок 1: Source

- Mathematically, a neural network is a function $f(w; x)$

Example: White-box Adversarial Attacks

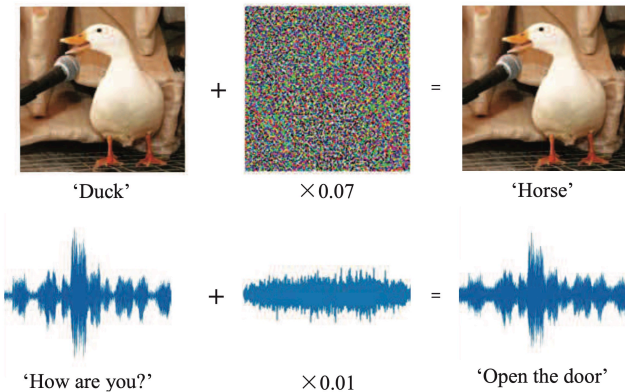


Рисунок 1: Source

- Mathematically, a neural network is a function $f(w; x)$
- Typically, input x is given and network weights w optimized

Example: White-box Adversarial Attacks

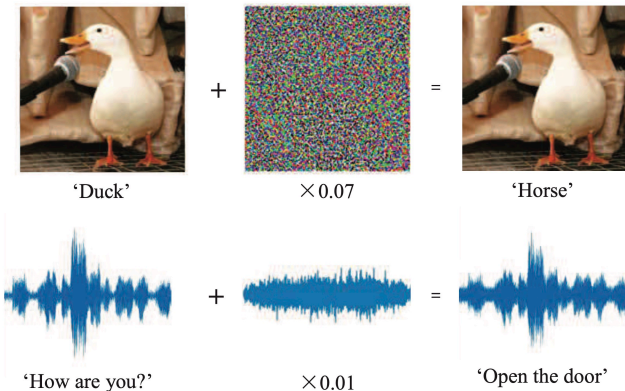


Рисунок 1: Source

- Mathematically, a neural network is a function $f(w; x)$
- Typically, input x is given and network weights w optimized
- Could also freeze weights w and optimize x , adversarially!

$$\min_{\delta} \text{size}(\delta) \quad \text{s.t.} \quad \text{pred}[f(w; x + \delta)] \neq y$$

or

$$\max_{\delta} l(w; x + \delta, y) \quad \text{s.t.} \quad \text{size}(\delta) \leq \epsilon, \quad 0 \leq x + \delta \leq 1$$

Idea of Projected Gradient Descent

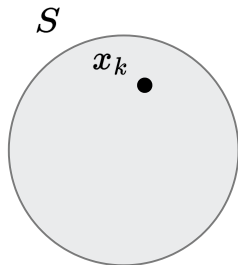


Рисунок 2: Suppose, we start from a point x_k .

Idea of Projected Gradient Descent

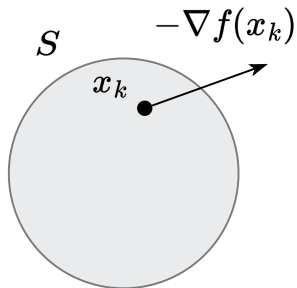


Рисунок 3: And go in the direction of $-\nabla f(x_k)$.

Idea of Projected Gradient Descent

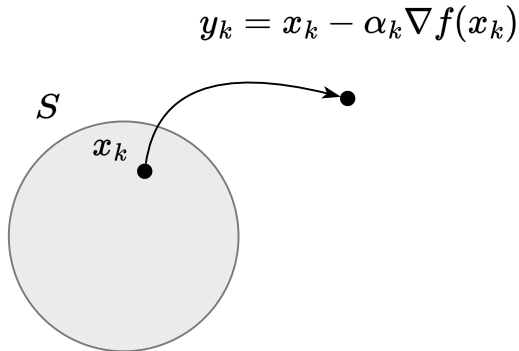


Рисунок 4: Occasionally, we can end up outside the feasible set.

Idea of Projected Gradient Descent

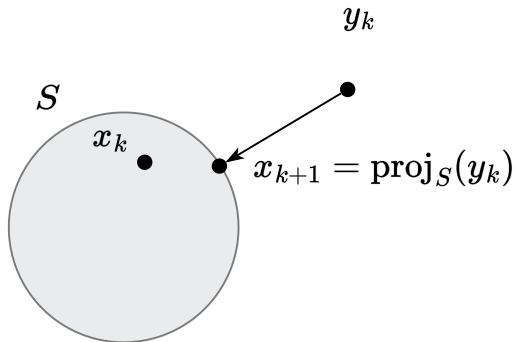


Рисунок 5: Solve this little problem with projection!

Idea of Projected Gradient Descent

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

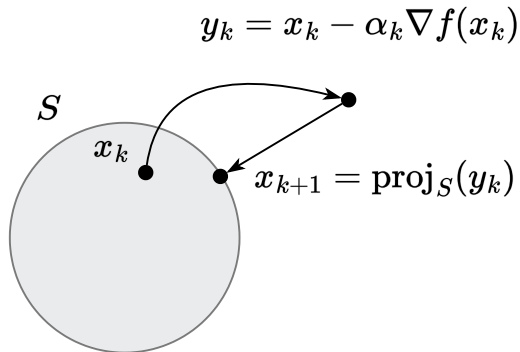


Рисунок 6: Illustration of Projected Gradient Descent algorithm

Projection

Projection

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

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

Projection

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

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

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 $\text{proj}_S(\mathbf{y}) \in S$:

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

Projection

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

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

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 $\text{proj}_S(\mathbf{y}) \in S$:

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

- **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.

Projection

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

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

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 $\text{proj}_S(\mathbf{y}) \in S$:

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

- **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.

Projection

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

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

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 $\text{proj}_S(\mathbf{y}) \in S$:

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

- **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 may not exist.

Projection

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

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

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 $\text{proj}_S(\mathbf{y}) \in S$:

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

- **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 may not exist.
- If a point is in set, then its projection is the point itself.

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

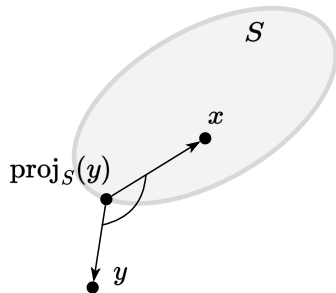


Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

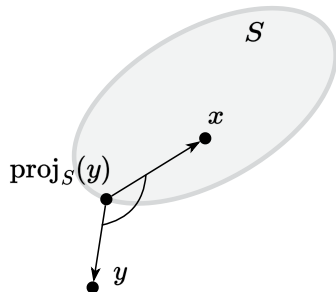


Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 (\text{proj}_S(y) - y)^T (x - \text{proj}_S(y)) \geq 0$$



Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 (\text{proj}_S(y) - y)^T (x - \text{proj}_S(y)) \geq 0$$

$$(y - \text{proj}_S(y))^T (x - \text{proj}_S(y)) \leq 0$$



Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 (\text{proj}_S(y) - y)^T (x - \text{proj}_S(y)) \geq 0$$

$$(y - \text{proj}_S(y))^T (x - \text{proj}_S(y)) \leq 0$$

2. Use cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ with $x = x - \text{proj}_S(y)$ and $y = y - \text{proj}_S(y)$. By the first property of the theorem:

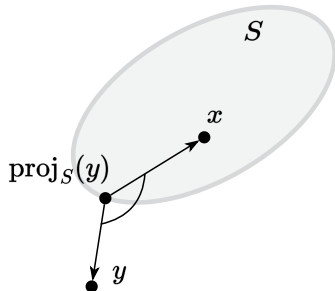


Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 (\text{proj}_S(y) - y)^T (x - \text{proj}_S(y)) \geq 0$$

$$(y - \text{proj}_S(y))^T (x - \text{proj}_S(y)) \leq 0$$

2. Use cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ with $x = x - \text{proj}_S(y)$ and $y = y - \text{proj}_S(y)$. By the first property of the theorem:

$$0 \geq 2x^T y = \|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 - \|x - y\|^2$$



Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

i Theorem

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

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

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

1. $\text{proj}_S(y)$ is minimizer of differentiable convex function $d(y, S, \|\cdot\|) = \|x - y\|^2$ over S . By first-order characterization of optimality.

$$\nabla d(\text{proj}_S(y))^T (x - \text{proj}_S(y)) \geq 0$$

$$2 (\text{proj}_S(y) - y)^T (x - \text{proj}_S(y)) \geq 0$$

$$(y - \text{proj}_S(y))^T (x - \text{proj}_S(y)) \leq 0$$

2. Use cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$ with $x = x - \text{proj}_S(y)$ and $y = y - \text{proj}_S(y)$. By the first property of the theorem:

$$0 \geq 2x^T y = \|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 - \|x - y\|^2$$

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

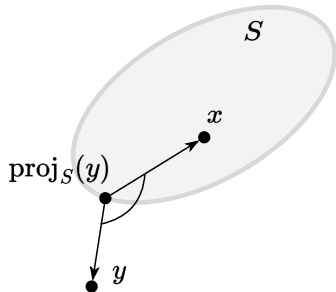


Рисунок 7: Obtuse or straight angle should be for any point $x \in S$

Projection operator is non-expansive

- A function f is called non-expansive if f is L -Lipschitz with $L \leq 1$ 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.

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

Projection operator is non-expansive

- A function f is called non-expansive if f is L -Lipschitz with $L \leq 1$ 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.

- Projection operator is non-expansive:

$$\|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

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

Projection operator is non-expansive

- A function f is called non-expansive if f is L -Lipschitz with $L \leq 1$ 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.

- Projection operator is non-expansive:

$$\|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

- Next: variational characterization implies non-expansiveness. i.e.,

$$\langle y - \text{proj}(y), x - \text{proj}(y) \rangle \leq 0 \quad \forall x \in S \quad \Rightarrow \quad \|\text{proj}(x) - \text{proj}(y)\|_2 \leq \|x - y\|_2.$$

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

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Replace x by $\pi(x)$ in Уравнение 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \quad (4)$$

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Replace x by $\pi(x)$ in Уравнение 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \quad (4)$$

Replace y by x and x by $\pi(y)$ in Уравнение 3

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \quad (5)$$

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Replace x by $\pi(x)$ in Уравнение 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \quad (4)$$

Replace y by x and x by $\pi(y)$ in Уравнение 3

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \quad (5)$$

(Уравнение 4)+(Уравнение 5) will cancel $\pi(y) - \pi(x)$, not good. So flip the sign of (Уравнение 5) gives

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \quad (6)$$

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Replace x by $\pi(x)$ in Уравнение 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \quad (4)$$

Replace y by x and x by $\pi(y)$ in Уравнение 3

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \quad (5)$$

(Уравнение 4)+(Уравнение 5) will cancel $\pi(y) - \pi(x)$, not good. So flip the sign of (Уравнение 5) gives

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \quad (6)$$

$$\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0$$

$$\langle y - x, \pi(x) - \pi(y) \rangle \leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y) \rangle$$

$$\langle y - x, \pi(y) - \pi(x) \rangle \geq \|\pi(x) - \pi(y)\|_2^2$$

$$\|(y - x)^\top (\pi(y) - \pi(x))\|_2 \geq \|\pi(x) - \pi(y)\|_2^2$$

Projection operator is non-expansive

Shorthand notation: let $\pi = \text{proj}$ and $\pi(x)$ denotes $\text{proj}(x)$.

Begins with the variational characterization / obtuse angle inequality

$$\langle y - \pi(y), x - \pi(y) \rangle \leq 0 \quad \forall x \in S. \quad (3)$$

Replace x by $\pi(x)$ in Уравнение 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \quad (4)$$

Replace y by x and x by $\pi(y)$ in Уравнение 3

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \quad (5)$$

(Уравнение 4)+(Уравнение 5) will cancel $\pi(y) - \pi(x)$, not good. So flip the sign of (Уравнение 5) gives

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \quad (6)$$

$$\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0$$

$$\langle y - x, \pi(x) - \pi(y) \rangle \leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y) \rangle$$

$$\langle y - x, \pi(y) - \pi(x) \rangle \geq \|\pi(x) - \pi(y)\|_2^2$$

$$\|(y - x)^\top (\pi(y) - \pi(x))\|_2 \geq \|\pi(x) - \pi(y)\|_2^2$$

By Cauchy-Schwarz inequality, the left-hand-side is upper bounded by

$\|y - x\|_2 \|\pi(y) - \pi(x)\|_2$, we get

$$\|y - x\|_2 \|\pi(y) - \pi(x)\|_2 \geq \|\pi(x) - \pi(y)\|_2^2.$$

Cancels $\|\pi(x) - \pi(y)\|_2$ finishes the proof.

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Build a hypothesis from the figure: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Build a hypothesis from the figure: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Check the inequality for a convex closed set: $(\pi - y)^T(x - \pi) \geq 0$

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Build a hypothesis from the figure: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Check the inequality for a convex closed set: $(\pi - y)^T(x - \pi) \geq 0$

$$\begin{aligned} & \left(x_0 - y + R \frac{y - x_0}{\|y - x_0\|} \right)^T \left(x - x_0 - R \frac{y - x_0}{\|y - x_0\|} \right) = \\ & \left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|} \right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|} \right) = \\ & \frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T ((x - x_0)\|y - x_0\| - R(y - x_0)) = \\ & \frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\| \right) = \\ & (R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R \right) \end{aligned}$$

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

Build a hypothesis from the figure: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Check the inequality for a convex closed set: $(\pi - y)^T(x - \pi) \geq 0$

The first factor is negative for point selection y .
The second factor is also negative, which follows from the Cauchy-Bunyakovsky inequality:

$$\begin{aligned} \left(x_0 - y + R \frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R \frac{y - x_0}{\|y - x_0\|}\right) &= \\ \left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) &= \\ \frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T ((x - x_0)\|y - x_0\| - R(y - x_0)) &= \\ \frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\|\right) &= \\ (R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R\right) \end{aligned}$$

Example: projection on the ball

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid \|x - x_0\| \leq R\}$, $y \notin S$

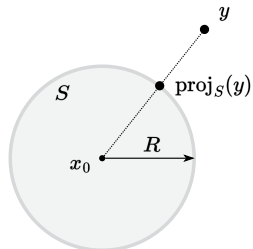
Build a hypothesis from the figure: $\pi = x_0 + R \cdot \frac{y - x_0}{\|y - x_0\|}$

Check the inequality for a convex closed set: $(\pi - y)^T(x - \pi) \geq 0$

$$\begin{aligned} \left(x_0 - y + R \frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R \frac{y - x_0}{\|y - x_0\|}\right) &= \\ \left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) &= \\ \frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T ((x - x_0)\|y - x_0\| - R(y - x_0)) &= \\ \frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\|\right) &= \\ (R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R\right) \end{aligned}$$

The first factor is negative for point selection y .
The second factor is also negative, which follows from the Cauchy-Bunyakovsky inequality:

$$\begin{aligned} (y - x_0)^T (x - x_0) &\leq \|y - x_0\| \|x - x_0\| \\ \frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R &\leq \frac{\|y - x_0\| \|x - x_0\|}{\|y - x_0\|} - R \end{aligned}$$



Example: projection on the halfspace

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Build a hypothesis from the figure: $\pi = y + \alpha c$. Coefficient α is chosen so that $\pi \in S$: $c^T \pi = b$, so:

Example: projection on the halfspace

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Build a hypothesis from the figure: $\pi = y + \alpha c$. Coefficient α is chosen so that $\pi \in S$: $c^T \pi = b$, so:

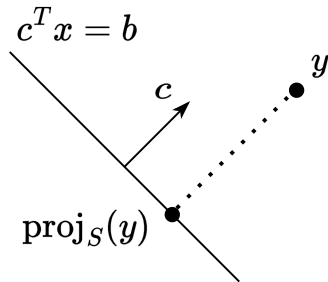


Рисунок 9: Hyperplane

Example: projection on the halfspace

Find $\pi_S(y) = \pi$, if $S = \{x \in \mathbb{R}^n \mid c^T x = b\}$, $y \notin S$. Build a hypothesis from the figure: $\pi = y + \alpha c$. Coefficient α is chosen so that $\pi \in S$: $c^T \pi = b$, so:

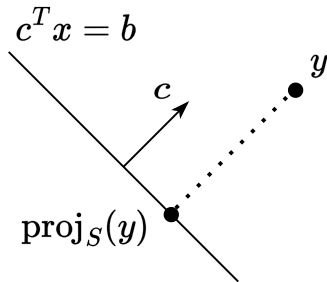


Рисунок 9: Hyperplane

$$c^T(y + \alpha c) = b$$

$$c^T y + \alpha c^T c = b$$

$$c^T y = b - \alpha c^T c$$

Check the inequality for a convex closed set:

$$(\pi - y)^T(x - \pi) \geq 0$$

$$(y + \alpha c - y)^T(x - y - \alpha c) =$$

$$\alpha c^T(x - y - \alpha c) =$$

$$\alpha(c^T x) - \alpha(c^T y) - \alpha^2(c^T c) =$$

$$\alpha b - \alpha(b - \alpha c^T c) - \alpha^2 c^T c =$$

$$\alpha b - \alpha b + \alpha^2 c^T c - \alpha^2 c^T c = 0 \geq 0$$

Projected Gradient Descent (PGD)

Idea

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



Рисунок 10: Illustration of Projected Gradient Descent algorithm



i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -smooth convex function. Then, for any $x, y \in \mathbb{R}^n$, the following inequality holds:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ or, equivalently,}$$
$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof

1. To prove this, we'll consider another function $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. It is obviously a convex function (as a sum of convex functions). And it is easy to verify, that it is an L -smooth function by definition, since $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ and $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.



i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -smooth convex function. Then, for any $x, y \in \mathbb{R}^n$, the following inequality holds:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ or, equivalently,}$$
$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof

1. To prove this, we'll consider another function $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. It is obviously a convex function (as a sum of convex functions). And it is easy to verify, that it is an L -smooth function by definition, since $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ and $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Now let's consider the smoothness parabolic property for the $\varphi(y)$ function:



i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -smooth convex function. Then, for any $x, y \in \mathbb{R}^n$, the following inequality holds:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ or, equivalently,}$$
$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof

1. To prove this, we'll consider another function $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. It is obviously a convex function (as a sum of convex functions). And it is easy to verify, that it is an L -smooth function by definition, since $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ and $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Now let's consider the smoothness parabolic property for the $\varphi(y)$ function:

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -smooth convex function. Then, for any $x, y \in \mathbb{R}^n$, the following inequality holds:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ or, equivalently,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof

1. To prove this, we'll consider another function $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. It is obviously a convex function (as a sum of convex functions). And it is easy to verify, that it is an L -smooth function by definition, since $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ and $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Now let's consider the smoothness parabolic property for the $\varphi(y)$ function:

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

$$\varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) \leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L} \nabla \varphi(y) \right\rangle + \frac{1}{2L} \|\nabla \varphi(y)\|_2^2$$

Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an L -smooth convex function. Then, for any $x, y \in \mathbb{R}^n$, the following inequality holds:

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y) \text{ or, equivalently,}$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 = \|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L (f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof

1. To prove this, we'll consider another function $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$. It is obviously a convex function (as a sum of convex functions). And it is easy to verify, that it is an L -smooth function by definition, since $\nabla \varphi(y) = \nabla f(y) - \nabla f(x)$ and $\|\nabla \varphi(y_1) - \nabla \varphi(y_2)\| = \|\nabla f(y_1) - \nabla f(y_2)\| \leq L\|y_1 - y_2\|$.
2. Now let's consider the smoothness parabolic property for the $\varphi(y)$ function:

$$\varphi(y) \leq \varphi(x) + \langle \nabla \varphi(x), y - x \rangle + \frac{L}{2} \|y - x\|_2^2$$

$$\begin{aligned} \varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) &\leq \varphi(y) + \left\langle \nabla \varphi(y), -\frac{1}{L} \nabla \varphi(y) \right\rangle + \frac{1}{2L} \|\nabla \varphi(y)\|_2^2 \\ \varphi\left(y - \frac{1}{L} \nabla \varphi(y)\right) &\leq \varphi(y) - \frac{1}{2L} \|\nabla \varphi(y)\|_2^2 \end{aligned}$$

Convergence tools

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

Convergence tools

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

Convergence tools

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle \nabla f(x), y \rangle$:

$$f(x) - \langle \nabla f(x), x \rangle \leq f(y) - \langle \nabla f(x), y \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle\nabla f(x), y - x\rangle)$$

Convergence tools

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle\nabla f(x), y - x\rangle)$$

switch x and y $\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle\nabla f(y), x - y\rangle)$

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle\nabla f(x), y - x\rangle)$$

switch x and y $\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle\nabla f(y), x - y\rangle)$

3. From the first order optimality conditions for the convex function $\nabla\varphi(y) = \nabla f(y) - \nabla f(x) = 0$. We can conclude, that for any x , the minimum of the function $\varphi(y)$ is at the point $y = x$. Therefore:

$$\varphi(x) \leq \varphi\left(y - \frac{1}{L}\nabla\varphi(y)\right) \leq \varphi(y) - \frac{1}{2L}\|\nabla\varphi(y)\|_2^2$$

4. Now, substitute $\varphi(y) = f(y) - \langle\nabla f(x), y\rangle$:

$$f(x) - \langle\nabla f(x), x\rangle \leq f(y) - \langle\nabla f(x), y\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

$$f(x) + \langle\nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_2^2 \leq f(y)$$

$$\|\nabla f(y) - \nabla f(x)\|_2^2 \leq 2L(f(y) - f(x) - \langle\nabla f(x), y - x\rangle)$$

switch x and y $\|\nabla f(x) - \nabla f(y)\|_2^2 \leq 2L(f(x) - f(y) - \langle\nabla f(y), x - y\rangle)$

The lemma has been proved. From the first view it does not make a lot of geometrical sense, but we will use it as a convenient tool to bound the difference between gradients.



i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable on \mathbb{R}^n . Then, the function f is μ -strongly convex if and only if for any $x, y \in \mathbb{R}^d$ the following holds:

$$\text{Strongly convex case } \mu > 0 \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

$$\text{Convex case } \mu = 0 \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$$

Proof

1. We will only give the proof for the strongly convex case, the convex one follows from it with setting $\mu = 0$. We start from necessity. For the strongly convex function

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

$$\text{sum} \quad \langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$f(x) - f(y) - \langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle$$

$$\langle \nabla f(y), x - y \rangle = \int_0^1 \langle \nabla f(y), x - y \rangle dt \quad \quad = \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \end{aligned}$$
$$\begin{aligned} &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), (x - y) \rangle dt \\ &= \int_0^1 t^{-1} \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \end{aligned}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt \end{aligned}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt \end{aligned}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Thus, we have a strong convexity criterion satisfied

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Thus, we have a strong convexity criterion satisfied

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2 \text{ or, equivalently:}$$

Convergence tools

2. For the sufficiency we assume, that $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2$. Using Newton-Leibniz theorem $f(x) = f(y) + \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt$:

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle dt - \langle \nabla f(y), x - y \rangle \\ \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y), x - y \rangle dt \\ y + t(x - y) - y &= t(x - y) \\ &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), t(x - y) \rangle dt \\ &\geq \int_0^1 t^{-1} \mu \|t(x - y)\|^2 dt = \mu \|x - y\|^2 \int_0^1 t dt = \frac{\mu}{2} \|x - y\|_2^2 \end{aligned}$$

Thus, we have a strong convexity criterion satisfied

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2} \|x - y\|_2^2 \text{ or, equivalently:}$$

$$\text{switch } x \text{ and } y \quad - \langle \nabla f(x), x - y \rangle \leq - \left(f(x) - f(y) + \frac{\mu}{2} \|x - y\|_2^2 \right)$$

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 $k > 0$:

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

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 $k > 0$:

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

1. Let's prove sufficient decrease lemma, assuming, that $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ and cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

(7)

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 $k > 0$:

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

1. Let's prove sufficient decrease lemma, assuming, that $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ and cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

$$\text{Smoothness: } f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

(7)

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 $k > 0$:

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

1. Let's prove sufficient decrease lemma, assuming, that $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ and cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

Smoothness:
$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Method:
$$= f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

(7)

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 $k > 0$:

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

1. Let's prove sufficient decrease lemma, assuming, that $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ and cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

$$\text{Smoothness: } f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$\text{Method: } = f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$\text{Cosine rule: } = f(x_k) - \frac{L}{2}(\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \quad (7)$$

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 $k > 0$:

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

1. Let's prove sufficient decrease lemma, assuming, that $y_k = x_k - \frac{1}{L}\nabla f(x_k)$ and cosine rule $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

Smoothness:
$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Method:
$$= f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

Cosine rule:
$$\begin{aligned} &= f(x_k) - \frac{L}{2}(\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2 \end{aligned} \quad (7)$$

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)$$

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)$$

3. We will use now projection property: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ with $x = x^*, y = y_k$:

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2$$
$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)$$

3. We will use now projection property: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ with $x = x^*, y = y_k$:

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2$$
$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

4. Now, using convexity and previous part:

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\begin{aligned}\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle &= \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right) \\ \langle \nabla f(x_k), x_k - x^* \rangle &= \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)\end{aligned}$$

3. We will use now projection property: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ with $x = x^*, y = y_k$:

$$\begin{aligned}\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 &\leq \|x^* - y_k\|^2 \\ \|y_k - x^*\|^2 &\geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2\end{aligned}$$

4. Now, using convexity and previous part:

Convexity:
$$f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$$

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\begin{aligned}\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle &= \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right) \\ \langle \nabla f(x_k), x_k - x^* \rangle &= \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)\end{aligned}$$

3. We will use now projection property: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ with $x = x^*, y = y_k$:

$$\begin{aligned}\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 &\leq \|x^* - y_k\|^2 \\ \|y_k - x^*\|^2 &\geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2\end{aligned}$$

4. Now, using convexity and previous part:

$$\begin{aligned}\text{Convexity:} \quad f(x_k) - f^* &\leq \langle \nabla f(x_k), x_k - x^* \rangle \\ &\leq \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|y_k - x_{k+1}\|^2 \right)\end{aligned}$$

Convergence rate for smooth and convex case

2. Now we do not immediately have progress at each step. Let's use again cosine rule:

$$\left\langle \frac{1}{L} \nabla f(x_k), x_k - x^* \right\rangle = \frac{1}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_k - x^* - \frac{1}{L} \nabla f(x_k)\|^2 \right)$$
$$\langle \nabla f(x_k), x_k - x^* \rangle = \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|y_k - x^*\|^2 \right)$$

3. We will use now projection property: $\|x - \text{proj}_S(y)\|^2 + \|y - \text{proj}_S(y)\|^2 \leq \|x - y\|^2$ with $x = x^*, y = y_k$:

$$\|x^* - \text{proj}_S(y_k)\|^2 + \|y_k - \text{proj}_S(y_k)\|^2 \leq \|x^* - y_k\|^2$$
$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

4. Now, using convexity and previous part:

Convexity:

$$f(x_k) - f^* \leq \langle \nabla f(x_k), x_k - x^* \rangle$$
$$\leq \frac{L}{2} \left(\frac{1}{L^2} \|\nabla f(x_k)\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2 - \|y_k - x_{k+1}\|^2 \right)$$

Sum for $i = 0, k-1$

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \sum_{i=0}^{k-1} \frac{1}{2L} \|\nabla f(x_i)\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

$$\begin{aligned}\sum_{i=0}^{k-1} [f(x_i) - f^*] &\leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2\end{aligned}$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

$$\begin{aligned}\sum_{i=0}^{k-1} [f(x_i) - f^*] &\leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \|x_0 - x^*\|^2\end{aligned}$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

$$\begin{aligned}\sum_{i=0}^{k-1} [f(x_i) - f^*] &\leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \|x_0 - x^*\|^2 \\ \sum_{i=0}^{k-1} f(x_i) - kf^* &\leq f(x_0) - f(x_k) + \frac{L}{2} \|x_0 - x^*\|^2\end{aligned}$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease inequality 7:

$$\begin{aligned}\sum_{i=0}^{k-1} [f(x_i) - f^*] &\leq \sum_{i=0}^{k-1} \left[f(x_i) - f(x_{i+1}) + \frac{L}{2} \|y_i - x_{i+1}\|^2 \right] + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 + \frac{L}{2} \|x_0 - x^*\|^2 - \frac{L}{2} \sum_{i=0}^{i-1} \|y_i - x_{i+1}\|^2 \\ &\leq f(x_0) - f(x_k) + \frac{L}{2} \|x_0 - x^*\|^2 \\ \sum_{i=0}^{k-1} f(x_i) - kf^* &\leq f(x_0) - f(x_k) + \frac{L}{2} \|x_0 - x^*\|^2 \\ \sum_{i=1}^k [f(x_i) - f^*] &\leq \frac{L}{2} \|x_0 - x^*\|^2\end{aligned}$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

we use the fact that $x_{k+1} = \text{proj}_S(y_k)$. By definition of projection,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

we use the fact that $x_{k+1} = \text{proj}_S(y_k)$. By definition of projection,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

and recall that $y_k = x_k - \frac{1}{L} \nabla f(x_k)$ implies $\|y_k - x_k\| = \frac{1}{L} \|\nabla f(x_k)\|$. Hence

$$\frac{L}{2} \|y_k - x_{k+1}\|^2 \leq \frac{L}{2} \|y_k - x_k\|^2 = \frac{L}{2} \frac{1}{L^2} \|\nabla f(x_k)\|^2 = \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

we use the fact that $x_{k+1} = \text{proj}_S(y_k)$. By definition of projection,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

and recall that $y_k = x_k - \frac{1}{L} \nabla f(x_k)$ implies $\|y_k - x_k\| = \frac{1}{L} \|\nabla f(x_k)\|$. Hence

$$\frac{L}{2} \|y_k - x_{k+1}\|^2 \leq \frac{L}{2} \|y_k - x_k\|^2 = \frac{L}{2} \frac{1}{L^2} \|\nabla f(x_k)\|^2 = \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

Substitute back into (*):

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{1}{2L} \|\nabla f(x_k)\|^2 = f(x_k).$$

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{L}{2} \|y_k - x_{k+1}\|^2,$$

we use the fact that $x_{k+1} = \text{proj}_S(y_k)$. By definition of projection,

$$\|y_k - x_{k+1}\| \leq \|y_k - x_k\|,$$

and recall that $y_k = x_k - \frac{1}{L} \nabla f(x_k)$ implies $\|y_k - x_k\| = \frac{1}{L} \|\nabla f(x_k)\|$. Hence

$$\frac{L}{2} \|y_k - x_{k+1}\|^2 \leq \frac{L}{2} \|y_k - x_k\|^2 = \frac{L}{2} \frac{1}{L^2} \|\nabla f(x_k)\|^2 = \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

Substitute back into (*):

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2 + \frac{1}{2L} \|\nabla f(x_k)\|^2 = f(x_k).$$

Hence

$$f(x_{k+1}) \leq f(x_k) \quad \text{for each } k,$$

so $\{f(x_k)\}$ is a monotonically nonincreasing sequence.

Convergence rate for smooth and convex case

7. Final convergence bound From step 5, we have already established

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.$$

Convergence rate for smooth and convex case

7. Final convergence bound From step 5, we have already established

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.$$

Convergence rate for smooth and convex case

7. Final convergence bound From step 5, we have already established

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.$$

Since $f(x_i)$ decreases in i , in particular $f(x_k) \leq f(x_i)$ for all $i \leq k$. Therefore

$$k [f(x_k) - f^*] \leq \sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2,$$

Convergence rate for smooth and convex case

7. Final convergence bound From step 5, we have already established

$$\sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2.$$

Since $f(x_i)$ decreases in i , in particular $f(x_k) \leq f(x_i)$ for all $i \leq k$. Therefore

$$k [f(x_k) - f^*] \leq \sum_{i=0}^{k-1} [f(x_i) - f^*] \leq \frac{L}{2} \|x_0 - x^*\|_2^2,$$

which immediately gives

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

This completes the proof of the $\mathcal{O}(\frac{1}{k})$ convergence rate for convex and L -smooth f under projection constraints.

Convergence rate for smooth strongly convex case

Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly convex. 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 $\alpha \leq \frac{1}{L}$ achieves the following convergence after iteration $k > 0$:

$$\|x_k - x^*\|_2^2 \leq (1 - \alpha\mu)^k \|x_0 - x^*\|_2^2$$

Proof

1. We first prove the stationary point property: $\text{proj}_S(x^* - \alpha \nabla f(x^*)) = x^*$.

This follows from the projection criterion and the first-order optimality condition for x^* . Let $y = x^* - \alpha \nabla f(x^*)$. We need to show $\langle y - x^*, x - x^* \rangle \leq 0$ for all $x \in S$.

$$\langle (x^* - \alpha \nabla f(x^*)) - x^*, x - x^* \rangle = -\alpha \langle \nabla f(x^*), x - x^* \rangle \leq 0$$

The inequality holds because $\alpha > 0$ and $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ is the optimality condition for x^* .

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\|x_{k+1} - x^*\|_2^2 = \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2$$

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2\end{aligned}$$

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{nonexpansiveness} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2\end{aligned}$$

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{nonexpansiveness} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\ &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2\end{aligned}$$

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{nonexpansiveness} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\ &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2\end{aligned}$$

2. Now we use smoothness from the convergence tools and strong convexity:

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{nonexpansiveness} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\ &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2\end{aligned}$$

2. Now we use smoothness from the convergence tools and strong convexity:

$$\text{smoothness } \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 \leq 2L(f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle)$$

Convergence rate for smooth strongly convex case

1. Considering the distance to the solution and using the stationary point property:

$$\begin{aligned}\|x_{k+1} - x^*\|_2^2 &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - x^*\|_2^2 \\ \text{stationary point property} &= \|\text{proj}_S(x_k - \alpha \nabla f(x_k)) - \text{proj}_S(x^* - \alpha \nabla f(x^*))\|_2^2 \\ \text{nonexpansiveness} &\leq \|x_k - \alpha \nabla f(x_k) - (x^* - \alpha \nabla f(x^*))\|_2^2 \\ &= \|x_k - x^*\|^2 - 2\alpha \langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle + \alpha^2 \|\nabla f(x_k) - \nabla f(x^*)\|_2^2\end{aligned}$$

2. Now we use smoothness from the convergence tools and strong convexity:

$$\begin{aligned}\text{smoothness} \quad \|\nabla f(x_k) - \nabla f(x^*)\|_2^2 &\leq 2L (f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle) \\ \text{strong convexity} \quad -\langle \nabla f(x_k) - \nabla f(x^*), x_k - x^* \rangle &\leq -\left(f(x_k) - f(x^*) + \frac{\mu}{2} \|x_k - x^*\|_2^2\right) - \langle \nabla f(x^*), x_k - x^* \rangle\end{aligned}$$

Convergence rate for smooth strongly convex case

3. Substitute it:

Convergence rate for smooth strongly convex case

3. Substitute it:

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

Convergence rate for smooth strongly convex case

3. Substitute it:

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

Convergence rate for smooth strongly convex case

3. Substitute it:

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

4. Due to convexity of f : $f(x_k) - f(x^*) - \langle \nabla f(x^*), x_k - x^* \rangle \geq 0$. Therefore, if we use $\alpha \leq \frac{1}{L}$:

$$\|x_{k+1} - x^*\|_2^2 \leq (1 - \alpha\mu) \|x_k - x^*\|^2,$$

which is exactly linear convergence of the method with up to $1 - \frac{\mu}{L}$ convergence rate.

Frank-Wolfe Method



Рисунок 11: Marguerite Straus Frank (1927-2024)

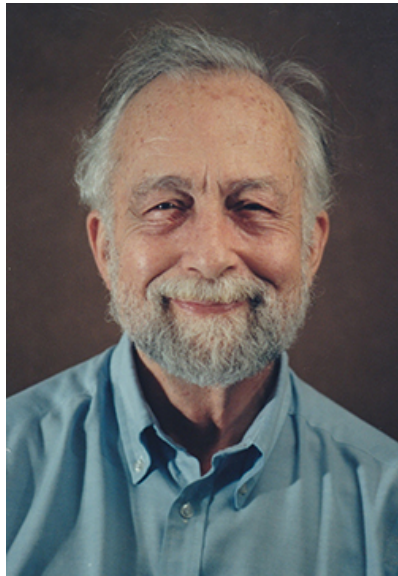


Рисунок 12: Philip Wolfe (1927-2016)

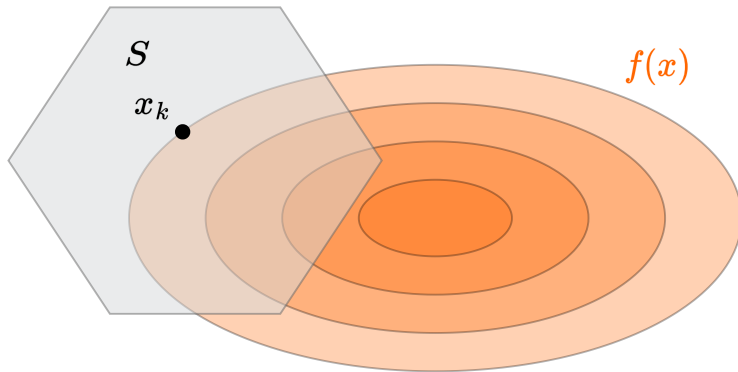


Рисунок 13: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

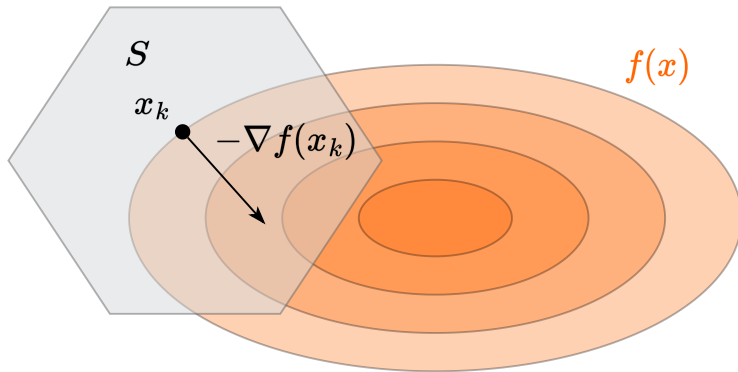


Рисунок 14: Illustration of Frank-Wolfe (conditional gradient) algorithm

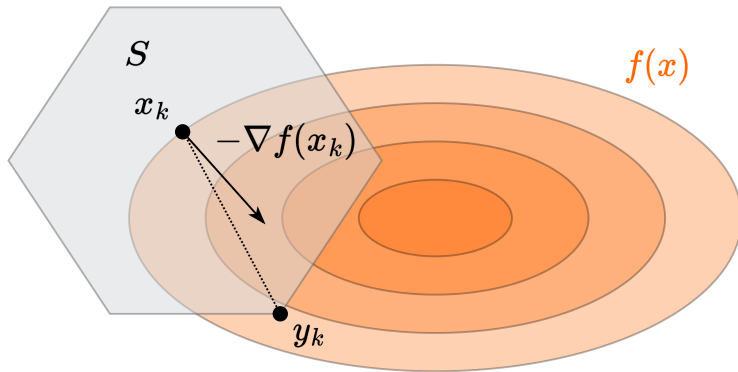


Рисунок 15: Illustration of Frank-Wolfe (conditional gradient) algorithm

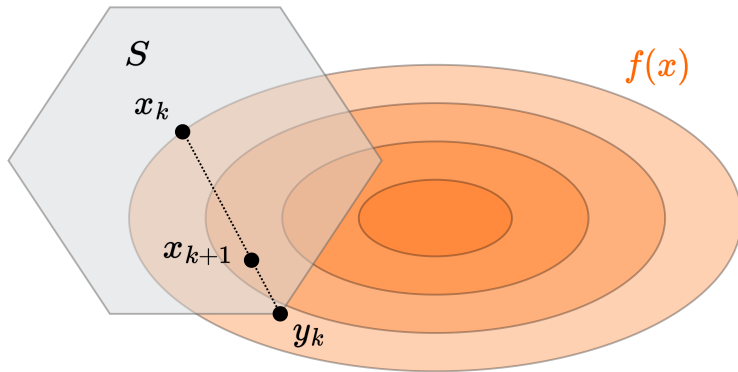


Рисунок 16: Illustration of Frank-Wolfe (conditional gradient) algorithm

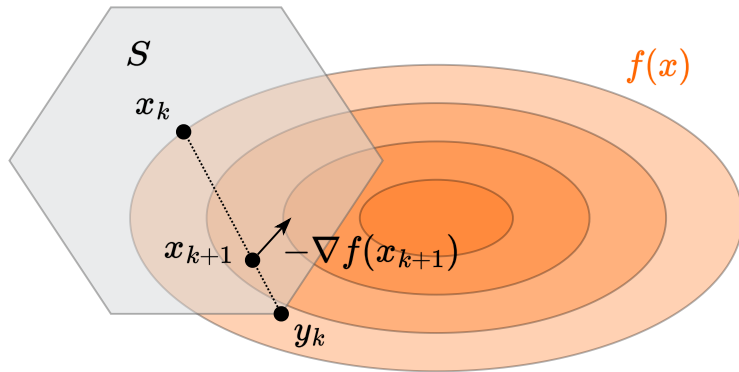


Рисунок 17: Illustration of Frank-Wolfe (conditional gradient) algorithm

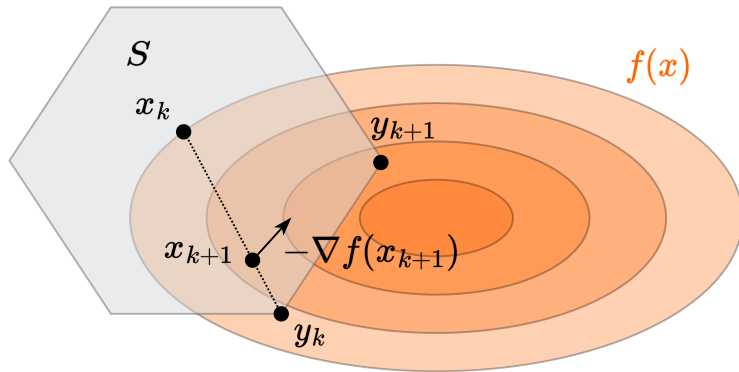


Рисунок 18: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

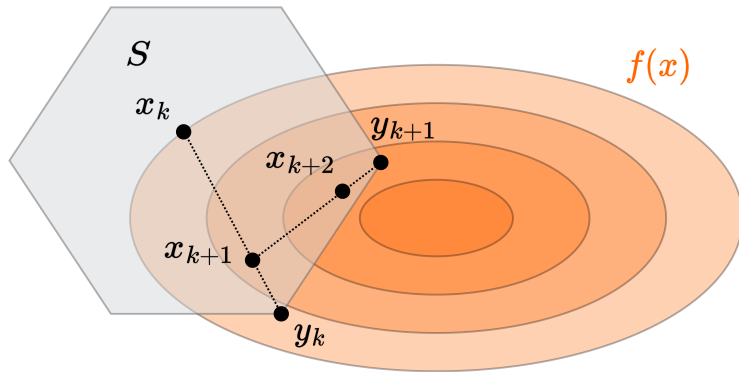


Рисунок 19: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

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

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

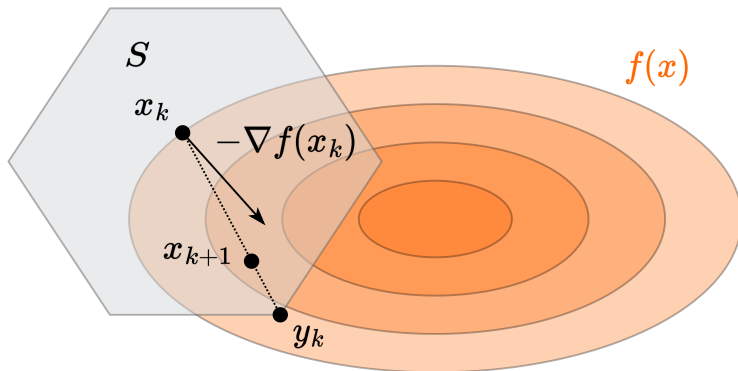


Рисунок 20: Illustration of Frank-Wolfe (conditional gradient) algorithm

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 Frank-Wolfe algorithm with step size $\gamma_k = \frac{k-1}{k+1}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

where $R = \max_{x,y \in S} \|x - y\|$ is the diameter of the set S .

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 Frank-Wolfe algorithm with step size $\gamma_k = \frac{k-1}{k+1}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

where $R = \max_{x,y \in S} \|x - y\|$ is the diameter of the set S .

1. By L -smoothness of f , we have:

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

Convergence rate for smooth and convex case

2. By convexity of f , for any $x \in S$, including x^* :

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

In particular, for $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

Convergence rate for smooth and convex case

2. By convexity of f , for any $x \in S$, including x^* :

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

In particular, for $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

3. By definition of y_k , we have $\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$, thus:

$$\langle \nabla f(x_k), y_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

Convergence rate for smooth and convex case

2. By convexity of f , for any $x \in S$, including x^* :

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

In particular, for $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

3. By definition of y_k , we have $\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$, thus:

$$\langle \nabla f(x_k), y_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

4. Combining the above inequalities:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq (1 - \gamma_k) \langle \nabla f(x_k), y_k - x_k \rangle + \frac{L(1 - \gamma_k)^2}{2} \|y_k - x_k\|^2 \\ &\leq (1 - \gamma_k) (f(x^*) - f(x_k)) + \frac{L(1 - \gamma_k)^2}{2} R^2 \end{aligned}$$

Convergence rate for smooth and convex case

2. By convexity of f , for any $x \in S$, including x^* :

$$\langle \nabla f(x_k), x - x_k \rangle \leq f(x) - f(x_k)$$

In particular, for $x = x^*$:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

3. By definition of y_k , we have $\langle \nabla f(x_k), y_k \rangle \leq \langle \nabla f(x_k), x^* \rangle$, thus:

$$\langle \nabla f(x_k), y_k - x_k \rangle \leq \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k)$$

4. Combining the above inequalities:

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq (1 - \gamma_k) \langle \nabla f(x_k), y_k - x_k \rangle + \frac{L(1 - \gamma_k)^2}{2} \|y_k - x_k\|^2 \\ &\leq (1 - \gamma_k) (f(x^*) - f(x_k)) + \frac{L(1 - \gamma_k)^2}{2} R^2 \end{aligned}$$

5. Rearranging terms:

$$f(x_{k+1}) - f(x^*) \leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2}$$

Convergence rate for smooth and convex case

6. Denoting $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$, we get:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}$$

Convergence rate for smooth and convex case

6. Denoting $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$, we get:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}$$

7. We will prove that $\delta_k \leq \frac{2}{k+1}$ by induction.

which gives us the desired result:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Convergence rate for smooth and convex case

6. Denoting $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$, we get:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}$$

7. We will prove that $\delta_k \leq \frac{2}{k+1}$ by induction.

- Base: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$

which gives us the desired result:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Convergence rate for smooth and convex case

6. Denoting $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$, we get:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}$$

7. We will prove that $\delta_k \leq \frac{2}{k+1}$ by induction.

- Base: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$
- Assume $\delta_k \leq \frac{2}{k+1}$

which gives us the desired result:


$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$

Convergence rate for smooth and convex case

6. Denoting $\delta_k = \frac{f(x_k) - f(x^*)}{LR^2}$, we get:

$$\delta_{k+1} \leq \gamma_k \delta_k + \frac{(1 - \gamma_k)^2}{2} = \frac{k-1}{k+1} \delta_k + \frac{2}{(k+1)^2}$$

7. We will prove that $\delta_k \leq \frac{2}{k+1}$ by induction.

- Base: $\delta_2 \leq \frac{1}{2} < \frac{2}{3}$
- Assume $\delta_k \leq \frac{2}{k+1}$
- Then $\delta_{k+1} \leq \frac{k-1}{k+1} \cdot \frac{2}{k+1} + \frac{2}{(k+1)^2} = \frac{2k}{k^2+2k+1} < \frac{2}{k+2}$ 

which gives us the desired result:

$$f(x_k) - f^* \leq \frac{2LR^2}{k+1}$$


Lower bound for Frank-Wolfe method ²

Theorem

Consider any algorithm that accesses the feasible set $S \subseteq \mathbb{R}^n$ only via a linear minimization oracle (LMO). Let the diameter of the set S be R . There exists an L -smooth strongly convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that this algorithm requires at least

$$\min \left(\frac{n}{2}, \frac{LR^2}{16\varepsilon} \right)$$

iterations (i.e., calls to the LMO) to construct a point $\hat{x} \in S$ with $f(\hat{x}) - \min_{x \in S} f(x) \leq \varepsilon$. The lower bound applies both for convex and strongly convex functions.

²  The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Lower bound for Frank-Wolfe method ²

Theorem

Consider any algorithm that accesses the feasible set $S \subseteq \mathbb{R}^n$ only via a linear minimization oracle (LMO). Let the diameter of the set S be R . There exists an L -smooth strongly convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that this algorithm requires at least

$$\min \left(\frac{n}{2}, \frac{LR^2}{16\varepsilon} \right)$$

iterations (i.e., calls to the LMO) to construct a point $\hat{x} \in S$ with $f(\hat{x}) - \min_{x \in S} f(x) \leq \varepsilon$. The lower bound applies both for convex and strongly convex functions.


Sketch of the proof. Consider the following optimization problem:

Note, that:

- f is 1-smooth;
- the diameter of S is $R = 2$;
- f is strongly convex.

$$\min_{x \in S} f(x) = \min_{x \in S} \frac{1}{2} \|x\|_2^2$$

$$S = \left\{ x \in \mathbb{R}^n \mid x \geq 0, \sum_{i=1}^n x_i = 1 \right\}$$

²  The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Lower bound for Frank-Wolfe method ³

1. The optimal solution is

$$x^* = \frac{1}{n} \mathbf{1} = \frac{1}{n} \sum_{i=1}^n e_i, \quad \text{and} \quad f(x^*) = \frac{1}{2n},$$

where $e_i = (0, \dots, 0, \underset{\text{position } i}{1}, 0, \dots, 0)^\top$ is the i -th standard basis vector.

Lower bound for Frank-Wolfe method ³

1. The optimal solution is

$$x^* = \frac{1}{n} \mathbf{1} = \frac{1}{n} \sum_{i=1}^n e_i, \quad \text{and} \quad f(x^*) = \frac{1}{2n},$$

where $e_i = (0, \dots, 0, \underset{\text{position } i}{1}, 0, \dots, 0)^\top$ is the i -th standard basis vector.

2. A linear minimization oracle (LMO) over S returns a vertex e_i . After k iterations, the method will have discovered at most k different basis vectors e_{i_1}, \dots, e_{i_k} . The best convex combination one can form is

$$\hat{x} = \frac{1}{k} \sum_{j=1}^k e_{i_j}.$$

Lower bound for Frank-Wolfe method ³

1. The optimal solution is

$$x^* = \frac{1}{n} \mathbf{1} = \frac{1}{n} \sum_{i=1}^n e_i, \quad \text{and} \quad f(x^*) = \frac{1}{2n},$$

where $e_i = (0, \dots, 0, \underset{\text{position } i}{1}, 0, \dots, 0)^\top$ is the i -th standard basis vector.

2. A linear minimization oracle (LMO) over S returns a vertex e_i . After k iterations, the method will have discovered at most k different basis vectors e_{i_1}, \dots, e_{i_k} . The best convex combination one can form is

$$\hat{x} = \frac{1}{k} \sum_{j=1}^k e_{i_j}.$$

3. Evaluating the function at \hat{x} , we obtain:

$$f(\hat{x}) - f(x^*) \geq \frac{1}{2} \left(\frac{1}{\min\{k, n\}} - \frac{1}{n} \right).$$

Lower bound for Frank-Wolfe method ³

1. The optimal solution is

$$x^* = \frac{1}{n} \mathbf{1} = \frac{1}{n} \sum_{i=1}^n e_i, \quad \text{and} \quad f(x^*) = \frac{1}{2n},$$

where $e_i = (0, \dots, 0, \underset{\text{position } i}{1}, 0, \dots, 0)^\top$ is the i -th standard basis vector.

2. A linear minimization oracle (LMO) over S returns a vertex e_i . After k iterations, the method will have discovered at most k different basis vectors e_{i_1}, \dots, e_{i_k} . The best convex combination one can form is


$$\hat{x} = \frac{1}{k} \sum_{j=1}^k e_{i_j}.$$

3. Evaluating the function at \hat{x} , we obtain:

$$f(\hat{x}) - f(x^*) \geq \frac{1}{2} \left(\frac{1}{\min\{k, n\}} - \frac{1}{n} \right).$$

4. To ensure that $f(\hat{x}) - f(x^*) \leq \varepsilon$, it is necessary that (full proof is in the paper):

$$k \geq \min \left\{ \frac{n}{2}, \frac{1}{4\varepsilon} \right\} = \min \left\{ \frac{n}{2}, \frac{LR^2}{16\varepsilon} \right\}.$$

³  The Complexity of Large-scale Convex Programming under a Linear Optimization Oracle

Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form


Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate.
This is the lower bound for LMO



Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected gradient descent, the rate is worse, but iteration could be cheaper and more sparse

Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected gradient descent, the rate is worse, but iteration could be cheaper and more sparse
- Recently, it was shown that for strongly convex sets, the rate can be improved to $O\left(\frac{1}{k^2}\right)$ ( paper)

Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected gradient descent, the rate is worse, but iteration could be cheaper and more sparse
- Recently, it was shown that for strongly convex sets, the rate can be improved to $O\left(\frac{1}{k^2}\right)$ ( paper)
- If we allow away steps, the convergence becomes linear ( paper) in strongly convex case

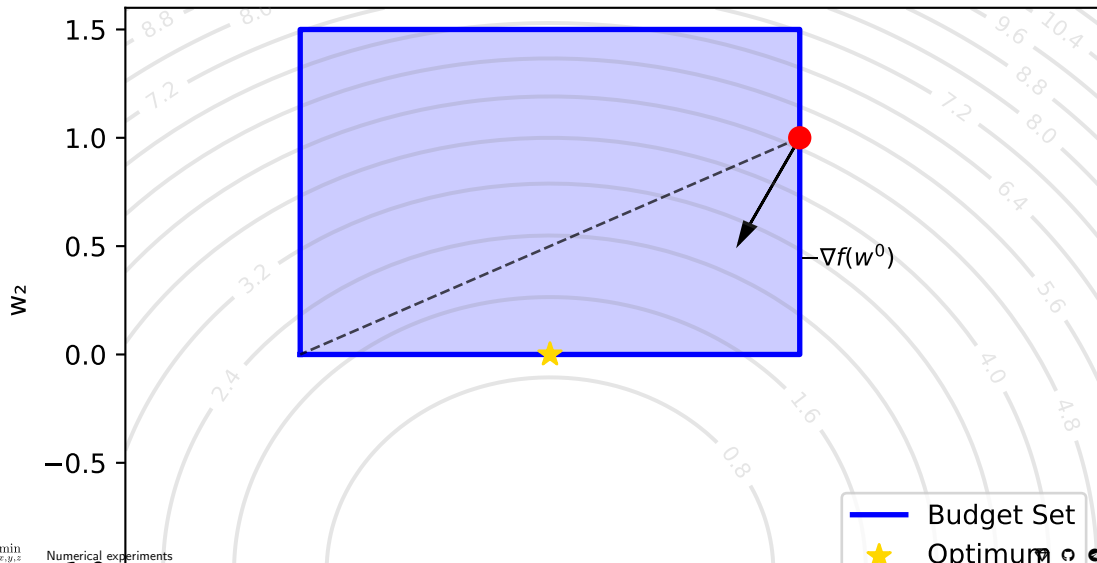
Frank-Wolfe method summary

- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected gradient descent, the rate is worse, but iteration could be cheaper and more sparse
- Recently, it was shown that for strongly convex sets, the rate can be improved to $O\left(\frac{1}{k^2}\right)$ (📄 paper)
- If we allow away steps, the convergence becomes linear (📄 paper) in strongly convex case
- Recent work showed the extension to non-smooth case (📄 paper) with convergence rate $O\left(\frac{1}{\sqrt{k}}\right)$

Numerical experiments

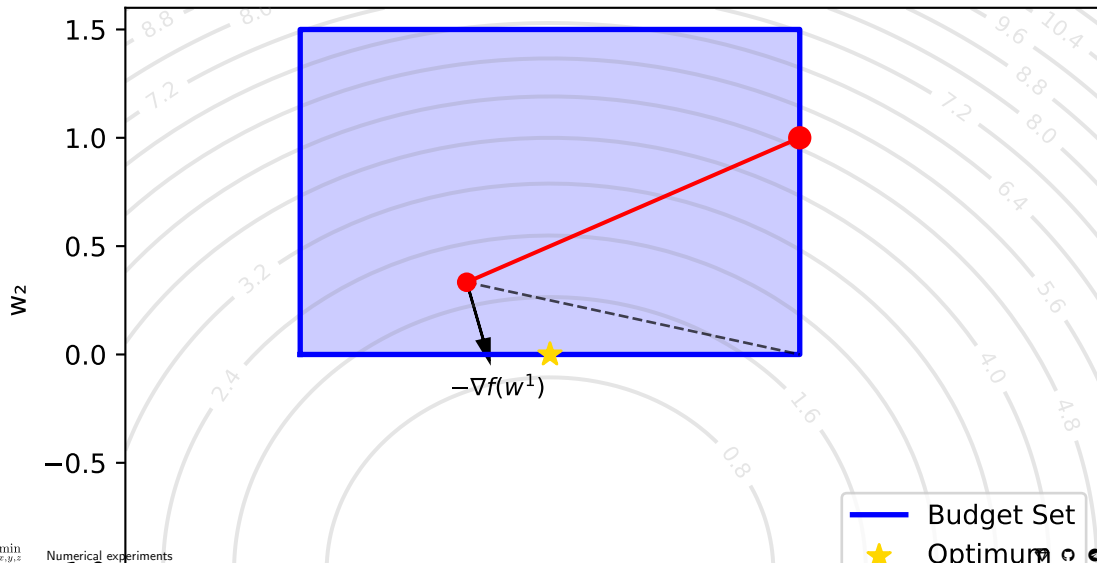
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 0



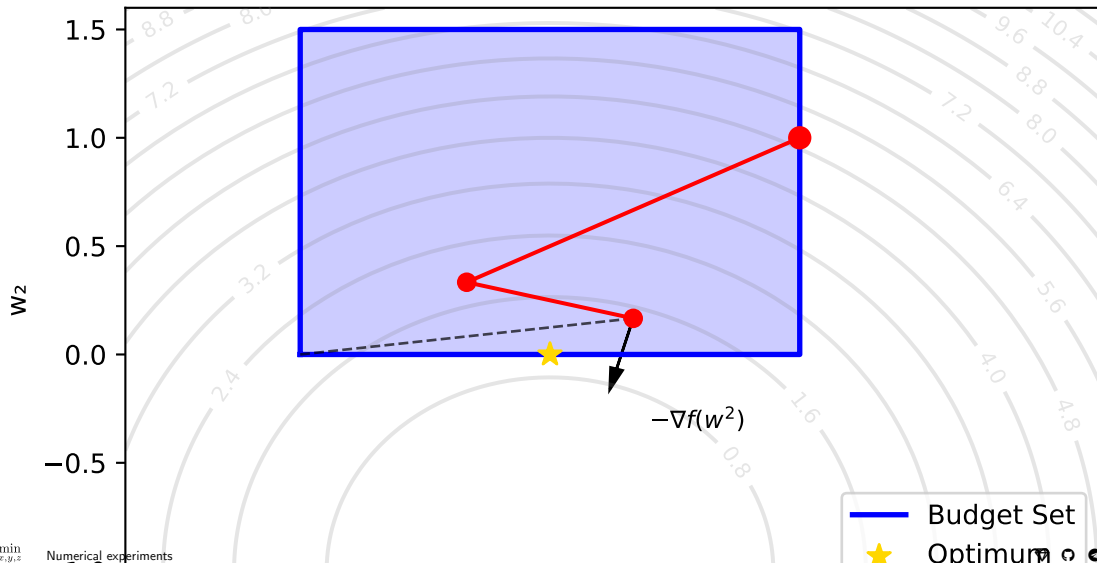
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 1



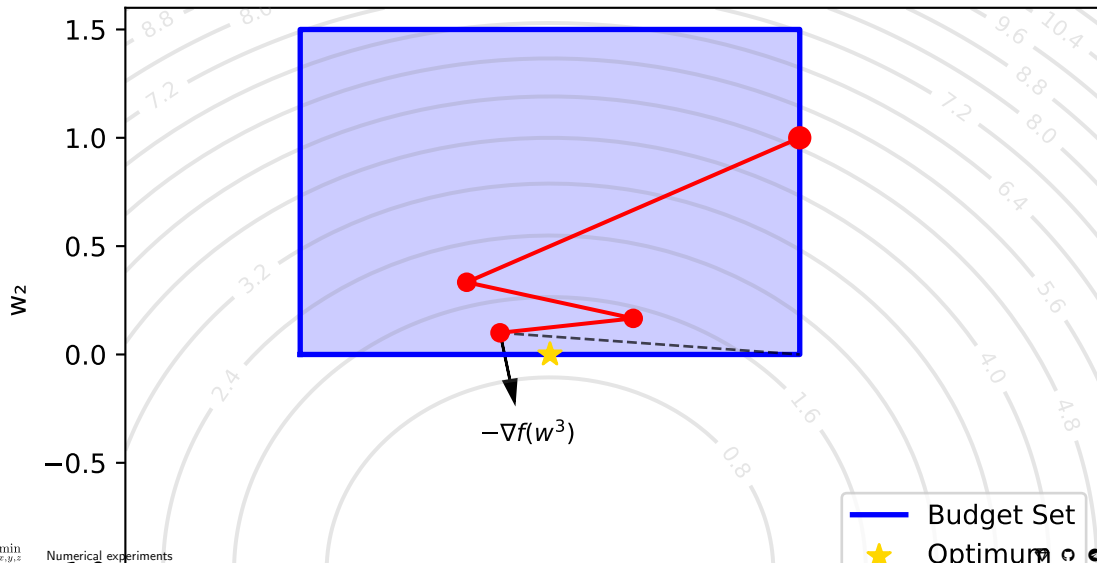
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 2



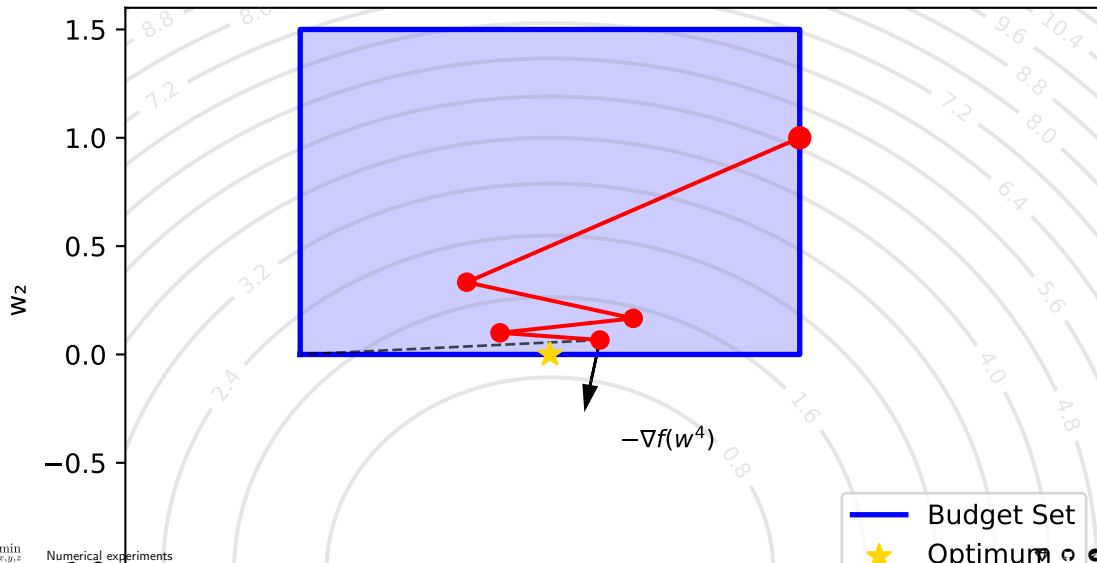
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 3



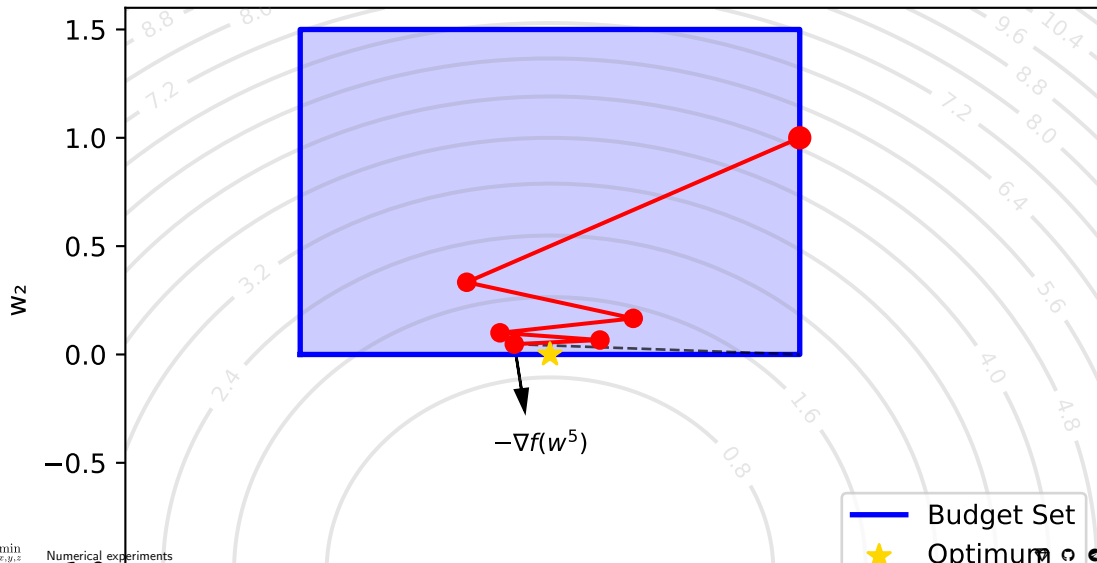
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 4



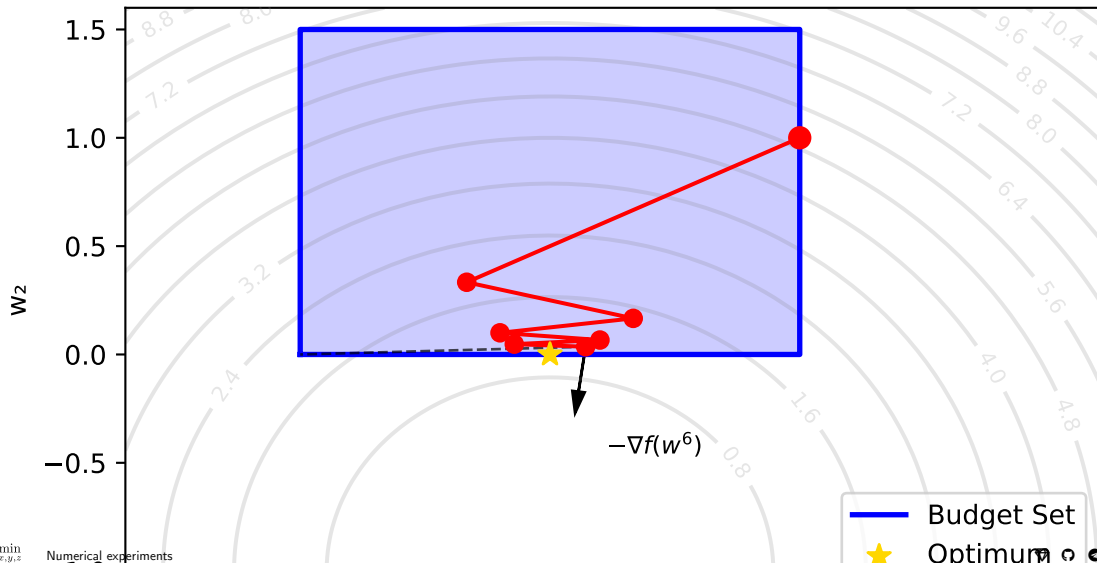
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 5



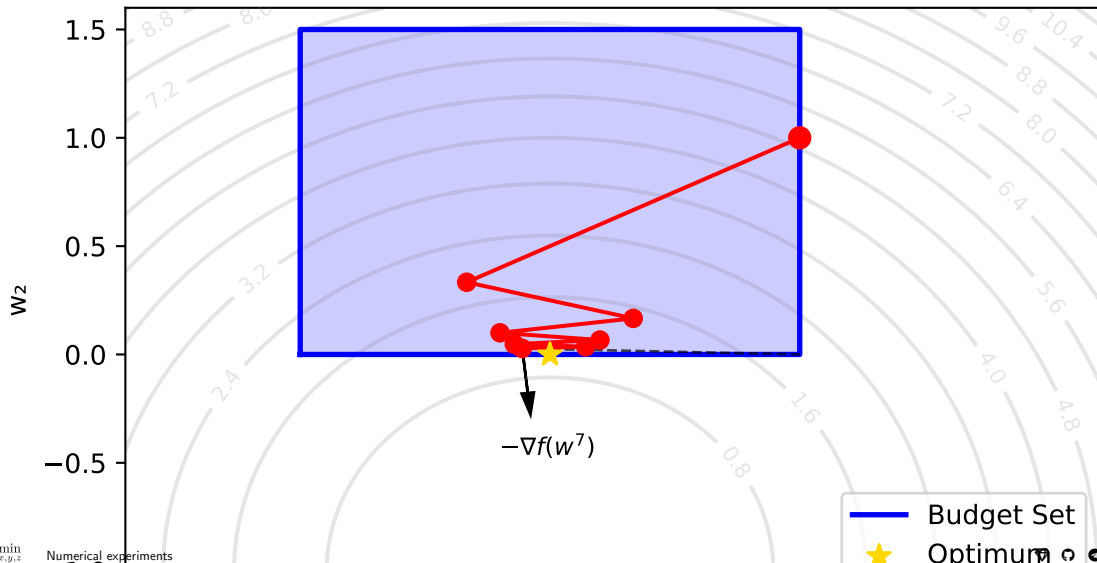
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 6



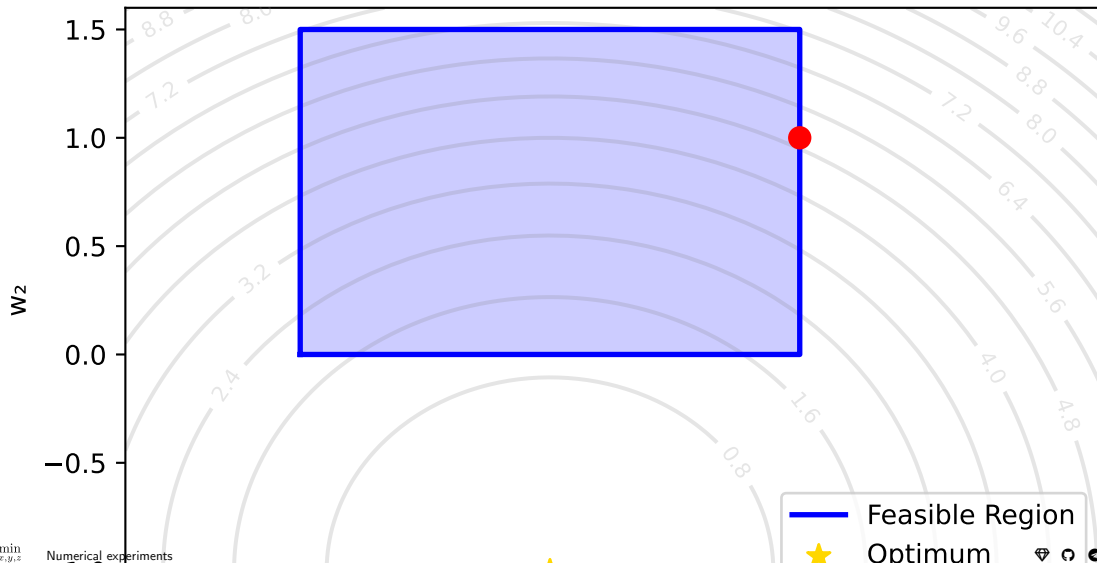
2d example. Frank-Wolfe method

Frank-Wolfe Method: Iteration 7



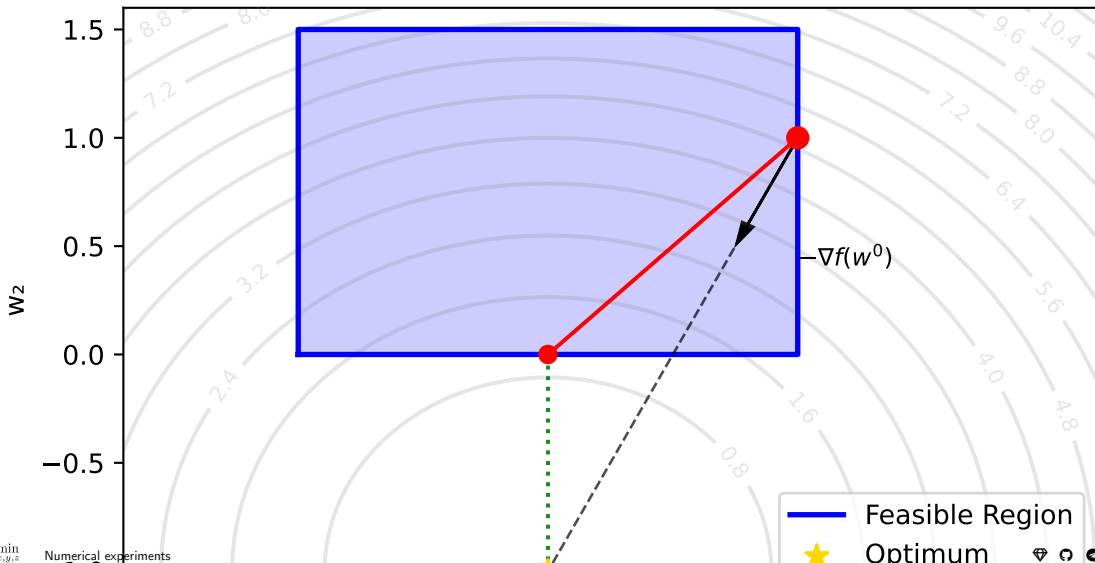
2d example. Projected gradient descent

Projected Gradient Descent: Iteration 0



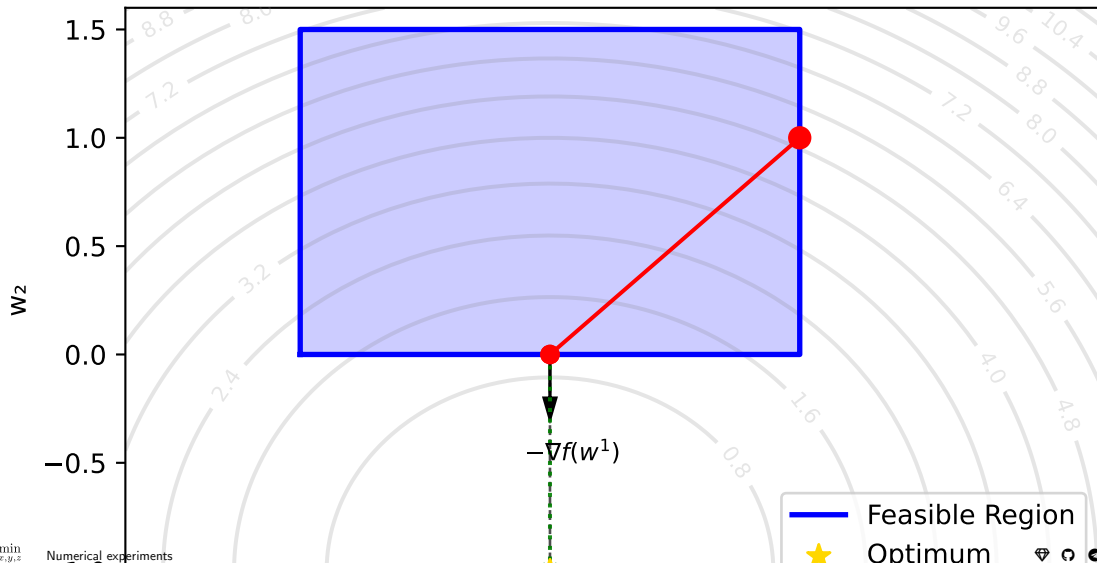
2d example. Projected gradient descent

Projected Gradient Descent: Iteration 1



2d example. Projected gradient descent

Projected Gradient Descent: Iteration 2



Quadratic function. Box constraints

$$\min_{\substack{x \in \mathbb{R}^n \\ -1 \preceq x \preceq 1}} \frac{1}{2} x^\top A x - b^\top x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].$$

The projection is simple:

$$\pi_S(x) = \text{clip}(x, -1, 1).$$

or

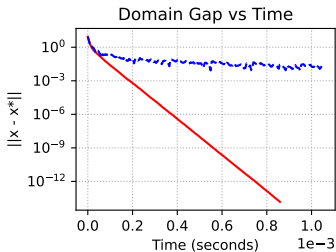
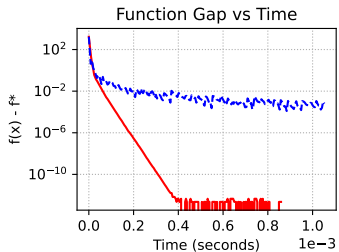
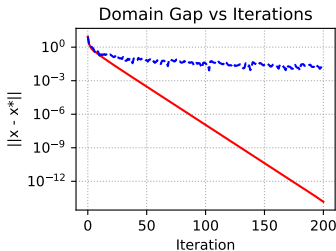
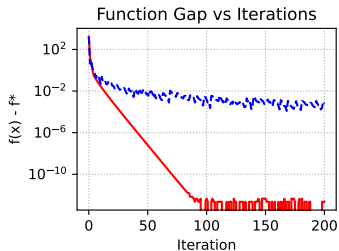
$$\pi_S(x) = \max(-1, \min(1, x)).$$

The linear minimization oracle (LMO) for a given gradient g is given by $y = \underset{z \in S}{\operatorname{argmin}} \langle g, z \rangle$.

Since the feasible set is separable across coordinates, the solution is computed coordinate-wise as

$$y_i = \begin{cases} -1, & \text{if } g_i > 0, \\ 1, & \text{if } g_i \leq 0. \end{cases}$$

Constrained convex quadratic problem: $n=80, \mu=0, L=10$



— Projected Gradient Descent - - - Frank-Wolfe

Quadratic function. Box constraints

$$\min_{\substack{x \in \mathbb{R}^n \\ -1 \preceq x \preceq 1}} \frac{1}{2} x^\top A x - b^\top x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [\mu; L].$$

The projection is simple:

$$\pi_S(x) = \text{clip}(x, -1, 1).$$

or

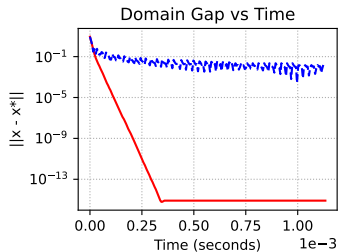
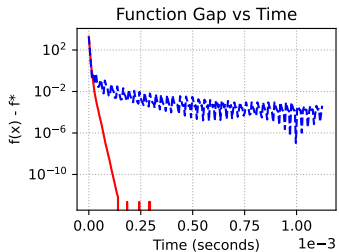
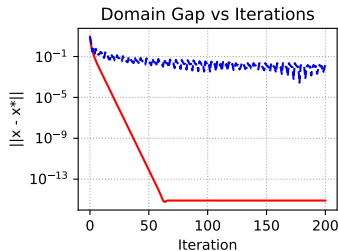
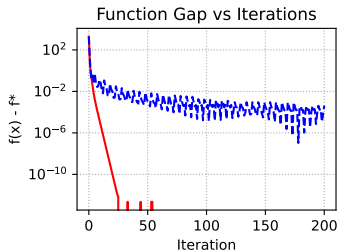
$$\pi_S(x) = \max(-1, \min(1, x)).$$

The linear minimization oracle (LMO) for a given gradient g is given by $y = \underset{z \in S}{\operatorname{argmin}} \langle g, z \rangle$.

Since the feasible set is separable across coordinates, the solution is computed coordinate-wise as

$$y_i = \begin{cases} -1, & \text{if } g_i > 0, \\ 1, & \text{if } g_i \leq 0. \end{cases}$$

Constrained strongly Convex quadratic problem: $n=80, \mu=1, L=10$



— Projected Gradient Descent - - - Frank-Wolfe

Quadratic function. Simplex constraints (Lucky problem with diagonal matrix)

$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, 1^T x = 1}} \frac{1}{2} x^T A x,$$

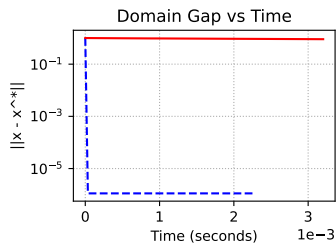
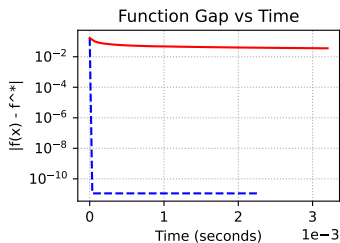
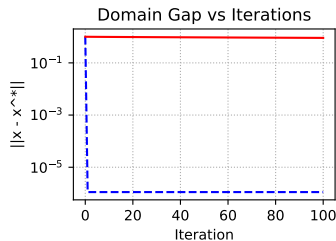
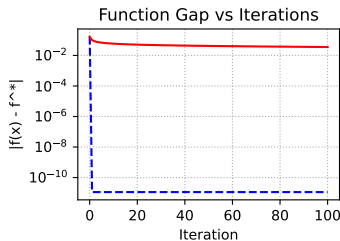
$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].$$

$$\min_{1^T x = 1, x \geq 0} \frac{1}{2} x^T A x, n = 200$$

Method	Update time, ms	LMO/Projection
PGD	0.0069	0.0167
FW	0.0070	0.0066

The projection onto the unit simplex $\pi_S(x)$ can be done in $\mathcal{O}(n \log n)$ or expected $\mathcal{O}(n)$ time.⁴ The LMO for a given gradient g is given by $y = \operatorname{argmin}_{z \in S} \langle g, z \rangle$. The solution corresponds to a vertex of the simplex:

$$y = e_j \quad \text{where} \quad j = \operatorname{argmin}_i g_i.$$



-- Frank-Wolfe — Projected Gradient Descent

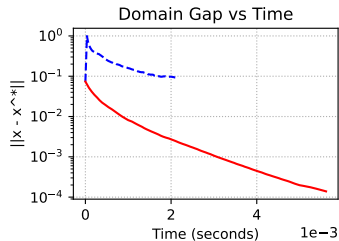
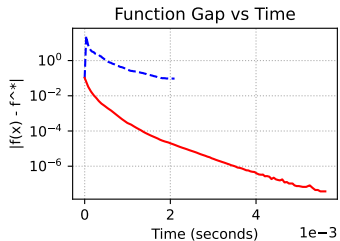
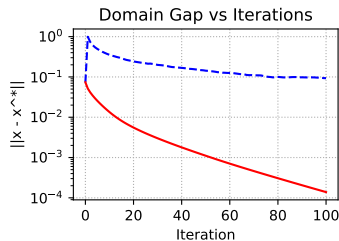
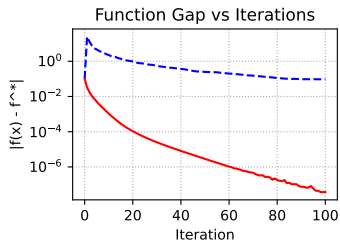
Quadratic function. Simplex constraints

$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, 1^T x = 1}} \frac{1}{2} x^T A x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].$$

$$\min_{1^T x = 1, x \geq 0} \frac{1}{2} x^T A x, \quad n = 200$$

Method	Update time, ms	LMO/Projection
PGD	0.0069	0.0420
FW	0.0069	0.0066



-- Frank-Wolfe — Projected Gradient Descent

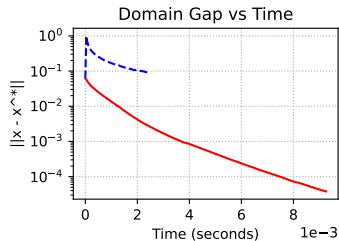
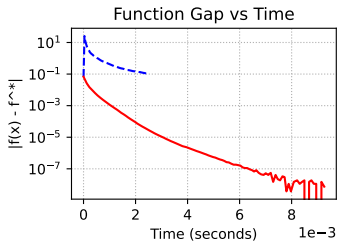
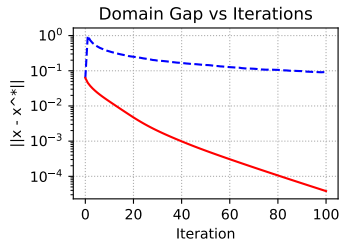
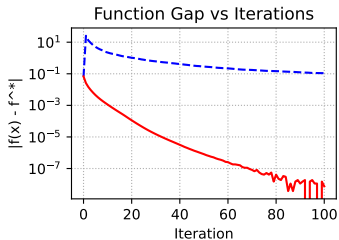
Quadratic function. Simplex constraints

$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, 1^T x = 1}} \frac{1}{2} x^T A x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [0; 100].$$

$$\min_{1^T x = 1, x \geq 0} \frac{1}{2} x^T A x, n = 300$$

Method	Update time, ms	LMO/Projection
PGD	0.0068	0.0761
FW	0.0069	0.0070



-- Frank-Wolfe — Projected Gradient Descent

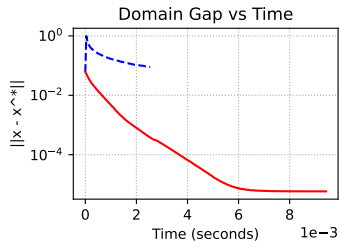
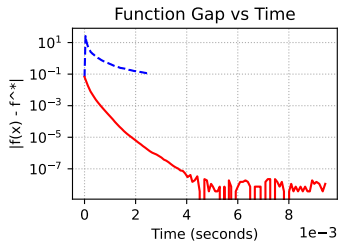
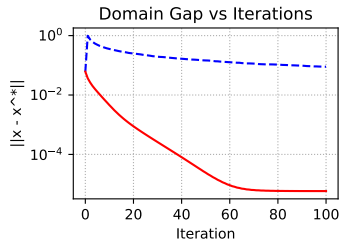
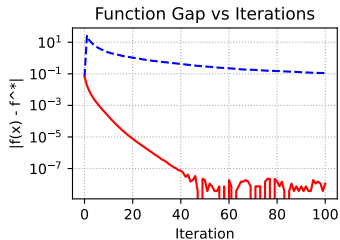
Quadratic function. Simplex constraints

$$\min_{\substack{x \in \mathbb{R}^n \\ x \geq 0, 1^T x = 1}} \frac{1}{2} x^T A x,$$

$$A \in \mathbb{R}^{n \times n}, \quad \lambda(A) \in [1; 100].$$

$$\min_{1^T x = 1, x \geq 0} \frac{1}{2} x^T A x, n = 300$$

Method	Update time, ms	LMO/Projection
PGD	0.0068	0.0752
FW	0.0067	0.0068



-- Frank-Wolfe — Projected Gradient Descent

PGD vs Frank-Wolfe

The key difference between PGD and FW is that PGD requires projection, while FW needs only linear minimization oracle (LMO).

In a recent book authors presented the following comparison table with complexities of linear minimizations and projections on some convex sets up to an additive error ϵ in the Euclidean norm.

Set	Linear minimization	Projection
n -dimensional ℓ_p -ball, $p \neq 1, 2, \infty$	$\mathcal{O}(n)$	$\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2}\right)$
Nuclear norm ball of $n \times m$ matrices	$\mathcal{O}\left(\nu \ln(m+n) \frac{\sqrt{\sigma_1}}{\sqrt{\epsilon}}\right)$	$\mathcal{O}(mn \min\{m, n\})$
Flow polytope on a graph with m vertices and n edges (capacity bound on edges)	$\mathcal{O}\left((n \log m)(n + m \log m)\right)$	$\tilde{\mathcal{O}}\left(\frac{n}{\epsilon^2}\right)$ or $\mathcal{O}(n^4 \log n)$
Birkhoff polytope ($n \times n$ doubly stochastic matrices)	$\mathcal{O}(n^3)$	$\tilde{\mathcal{O}}\left(\frac{n^2}{\epsilon^2}\right)$

When ϵ is missing, there is no additive error. The $\tilde{\mathcal{O}}$ hides polylogarithmic factors in the dimensions and polynomial factors in constants related to the distance to the optimum. For the nuclear norm ball, i.e., the spectralhedron, ν denotes the number of non-zero entries and σ_1 denotes the top singular value of the projected matrix.