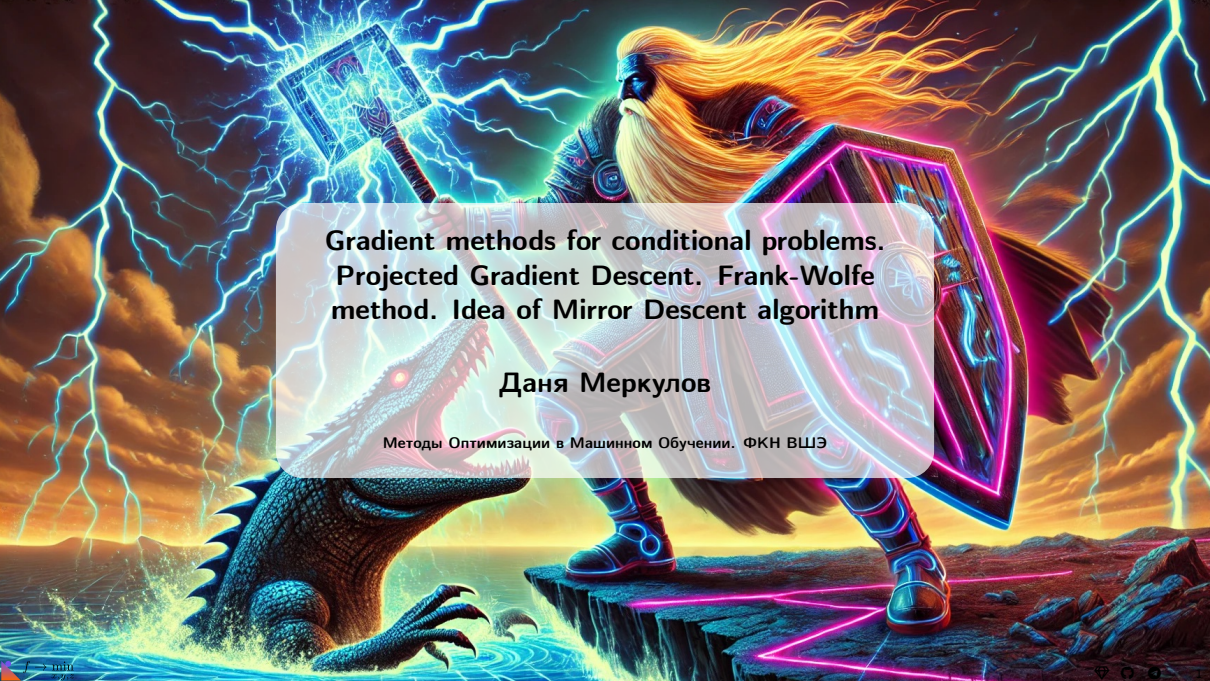


Gradient methods for conditional problems. Projected Gradient Descent. Frank-Wolfe method. Idea of Mirror Descent algorithm

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:

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

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:

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

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:

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

Gradient Descent is a great way to solve unconstrained problem

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

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:

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

Gradient Descent is a great way to solve unconstrained problem

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

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

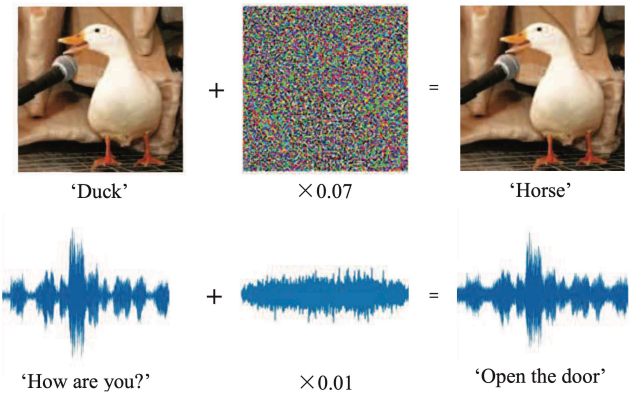


Figure 1: Source

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

Example: White-box Adversarial Attacks

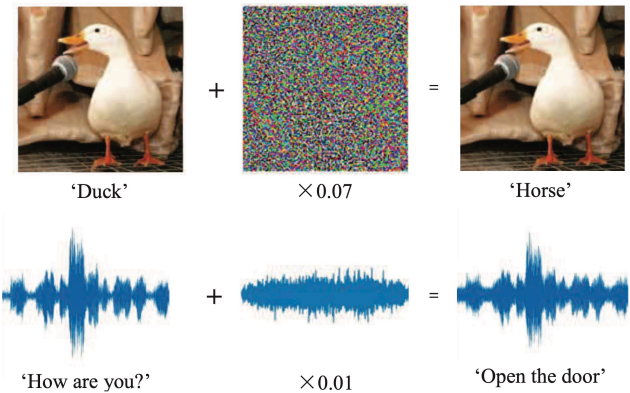


Figure 1: Source

- 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

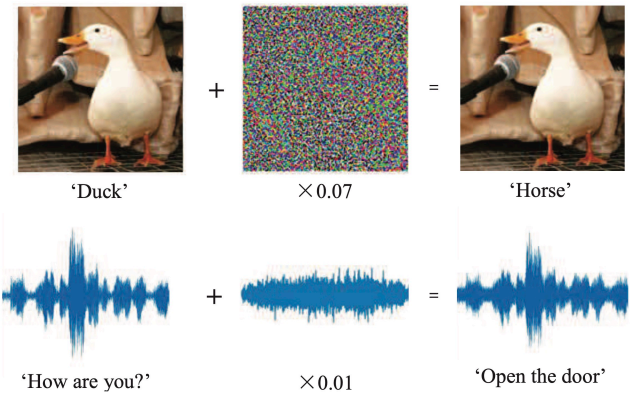


Figure 1: Source

- 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

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

Idea of Projected Gradient Descent

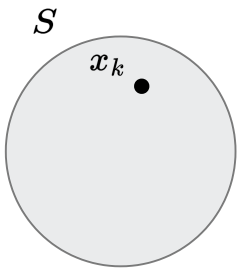


Figure 2: Suppose, we start from a point x_k.

Idea of Projected Gradient Descent

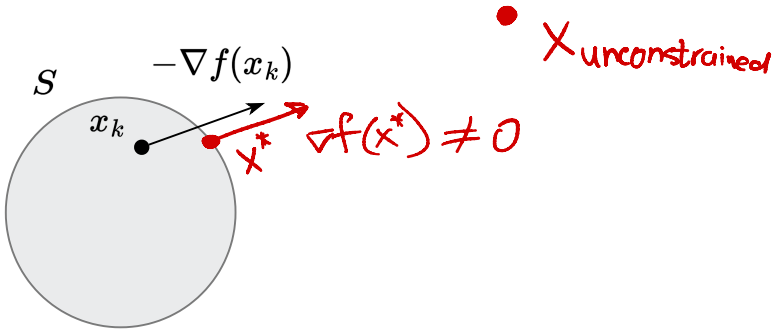


Figure 3: And go in the direction of $-\nabla f(x_k)$.

Idea of Projected Gradient Descent

$$y_k = x_k - \alpha_k \nabla f(x_k)$$

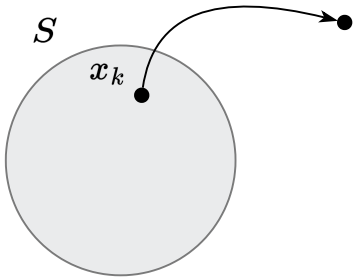


Figure 4: Occasionally, we can end up outside the feasible set.

Idea of Projected Gradient Descent

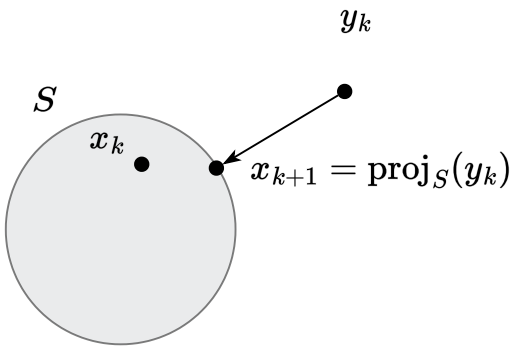


Figure 5: Solve this little problem with projection!

Idea of Projected Gradient Descent

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

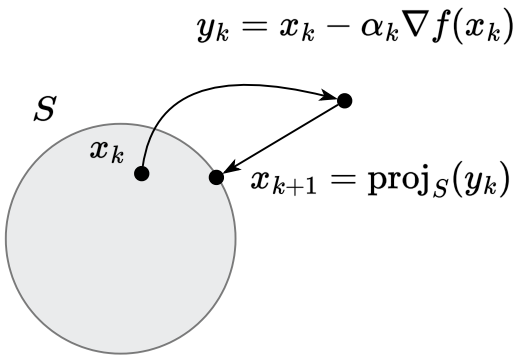


Figure 6: Illustration of Projected Gradient Descent algorithm

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

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

если это не так, см. зеркальный спуск Mirror Descent

Projection

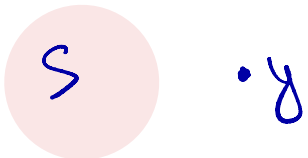
The distance d from point $\mathbf{y} \in \mathbb{R}^n$ to closed set $S \subset \mathbb{R}^n$:

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

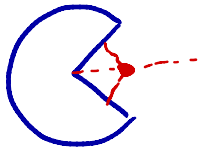
We will focus on Euclidean projection (other options are possible) of a point $\mathbf{y} \in \mathbb{R}^n$ on set $S \subseteq \mathbb{R}^n$ is a point $\text{proj}_S(\mathbf{y}) \in S$:

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

- **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.



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

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

- **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.

Projection

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

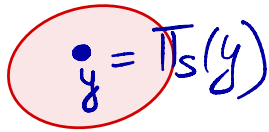
$$d(\mathbf{y}, S, \|\cdot\|) = \inf\{\|x - y\| \mid x \in S\}$$

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

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

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

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

- **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)

Theorem

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

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

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

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.

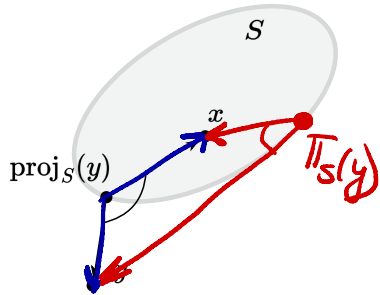


Figure 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

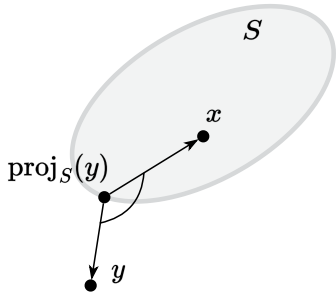


Figure 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

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

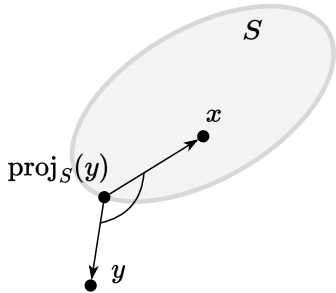


Figure 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

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

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

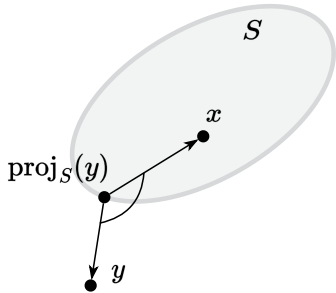


Figure 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

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

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

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:

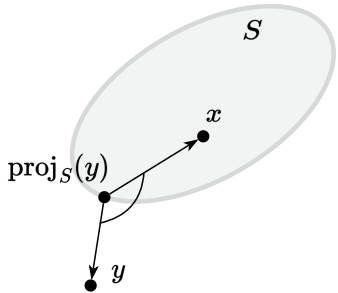


Figure 7: Obtuse or straight angle should be for any point $x \in S$

$x^T y = \|x\| \cdot \|y\| \cdot \cos\alpha$

$\cos\alpha = \dfrac{\|x\|^2 + \|y\|^2 - \|x-y\|^2}{\|x\| \cdot \|y\|}$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

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

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

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:

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

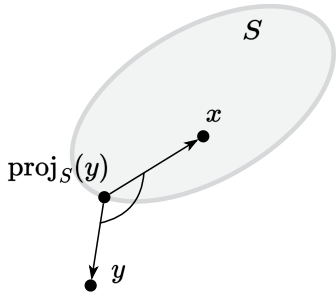


Figure 7: Obtuse or straight angle should be for any point $x \in S$

Projection criterion (Bourbaki-Cheney-Goldstein inequality)

Theorem

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

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

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

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.

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

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

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

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:

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

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

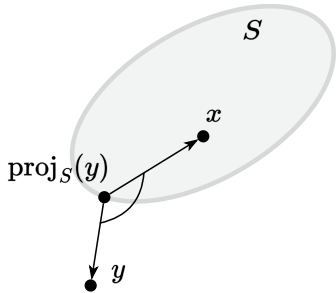


Figure 7: Obtuse or straight angle should be for any point $x \in S$

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 $x, y \in \mathrm{dom} f$,

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

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

$\forall x, y \quad \|\Pi(x) - \Pi(y)\| \leq \|x - y\|$

[1] Non-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 $x, y \in \mathrm{dom} f$,

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

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

- Projection operator is non-expansive:

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

[1] Non-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 $x, y \in \mathrm{dom} f$,

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

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

- Projection operator is non-expansive:

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

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

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

[1] Non-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

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

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

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

$x = \pi(x)$

Replace x by $\pi(x)$ in Equation 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

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

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

Replace x by $\pi(x)$ in Equation 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

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

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \tag{5}$$

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

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

Replace x by $\pi(x)$ in Equation 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

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

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \tag{5}$$

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

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \tag{6}$$

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

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

Replace x by $\pi(x)$ in Equation 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

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

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \tag{5}$$

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

$$\boxed{\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0.} \tag{6}$$

$$\begin{aligned}
\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y) \rangle &\leq 0 \\
\langle y - x, \pi(x) - \pi(y) \rangle &\leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y) \rangle \\
\langle y - x, \pi(y) - \pi(x) \rangle &\geq \|\pi(x) - \pi(y)\|_2^2 \\
\|(y - x)^\top (\pi(y) - \pi(x))\|_2 &\geq \|\pi(x) - \pi(y)\|_2^2
\end{aligned}$$

Projection operator is non-expansive

$\forall x, y \in \mathbb{R}^n : \|\Pi_S(x) - \Pi_S(y)\| \leq \|x - y\|$

S-ben. zárvány.

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

Begins with the variational characterization / obtuse angle inequality

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

Replace x by $\pi(x)$ in Equation 3

$$\langle y - \pi(y), \pi(x) - \pi(y) \rangle \leq 0. \tag{4}$$

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

$$\langle x - \pi(x), \pi(y) - \pi(x) \rangle \leq 0. \tag{5}$$

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

$$\langle \pi(x) - x, \pi(x) - \pi(y) \rangle \leq 0. \tag{6}$$

$$\begin{aligned}
\langle y - \pi(y) + \pi(x) - x, \pi(x) - \pi(y) \rangle &\leq 0 \\
\langle y - x, \pi(x) - \pi(y) \rangle &\leq -\langle \pi(x) - \pi(y), \pi(x) - \pi(y) \rangle \\
\langle y - x, \pi(y) - \pi(x) \rangle &\geq \|\pi(x) - \pi(y)\|_2^2 \\
\|(y - x)^\top (\pi(y) - \pi(x))\|_2 &\geq \|\pi(x) - \pi(y)\|_2^2
\end{aligned}$$

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

$$\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|}\right) =$$

$$\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) =$$

$$\frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T \left((x - x_0)\|y - x_0\| - R(y - x_0)\right) =$$

$$\frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\|\right) =$$

$$(R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R\right)$$

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

$$\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|}\right) =$$

$$\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) =$$

$$\frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T \left((x - x_0)\|y - x_0\| - R(y - x_0)\right) =$$

$$\frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\|\right) =$$

$$(R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R\right)$$

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

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

$$\left(x_0 - y + R\frac{y - x_0}{\|y - x_0\|}\right)^T \left(x - x_0 - R\frac{y - x_0}{\|y - x_0\|}\right) =$$

$$\left(\frac{(y - x_0)(R - \|y - x_0\|)}{\|y - x_0\|}\right)^T \left(\frac{(x - x_0)\|y - x_0\| - R(y - x_0)}{\|y - x_0\|}\right) =$$

$$\frac{R - \|y - x_0\|}{\|y - x_0\|^2} (y - x_0)^T \left((x - x_0)\|y - x_0\| - R(y - x_0)\right) =$$

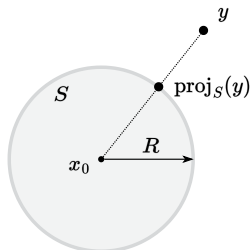
$$\frac{R - \|y - x_0\|}{\|y - x_0\|} \left((y - x_0)^T (x - x_0) - R\|y - x_0\|\right) =$$

$$(R - \|y - x_0\|) \left(\frac{(y - x_0)^T (x - x_0)}{\|y - x_0\|} - R\right)$$

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

$$(y - x_0)^T (x - x_0) \leq \|y - x_0\| \|x - x_0\|$$

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



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:

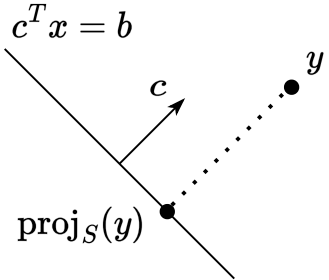


Figure 9: Hyperplane

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:

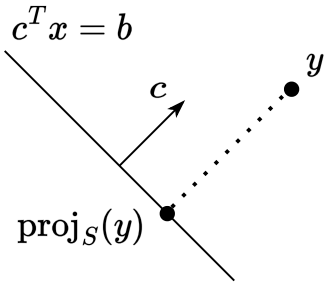


Figure 9: Hyperplane

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

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

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

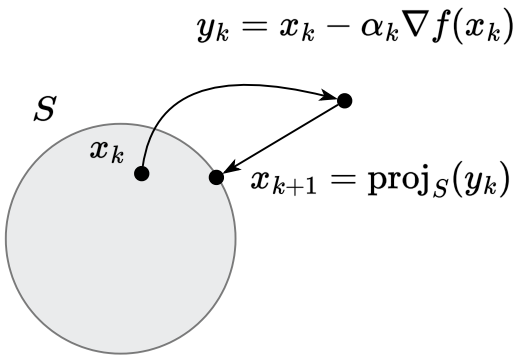


Figure 10: Illustration of Projected Gradient Descent algorithm

Convergence rate for smooth and convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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 $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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 $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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: $f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2$

(7)

Convergence rate for smooth and convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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 $k > 0$:

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

$y_k = x_k - \frac{1}{L}\nabla f(x_k)$

$x_{k+1} = \Pi_S(y_k)$

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

$$\begin{aligned} \text{Smoothness:} \quad & f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ \text{Method:} \quad & = f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \end{aligned} \tag{7}$$

Convergence rate for smooth and convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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 $2x^T y = \|x\|^2 + \|y\|^2 - \|x - y\|^2$:

$$\begin{aligned}
\text{Smoothness:} \quad & f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
\text{Method:} \quad & = f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
\text{Cosine rule:} \quad & = f(x_k) - \frac{L}{2}\left(\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2\right) + \frac{L}{2}\|x_{k+1} - x_k\|^2
\end{aligned} \tag{7}$$

Convergence rate for smooth and convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. Let $S \subseteq \mathbb{R}^n$d 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 $k > 0$:

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

$y_k = x_k - \frac{1}{L}\nabla f(x_k)$

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

$$\begin{aligned}
\text{Smoothness:} \quad & f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
\text{Method:} \quad & = f(x_k) - L\langle y_k - x_k, x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
\text{Cosine rule:} \quad & = f(x_k) - \frac{L}{2}(\|y_k - x_k\|^2 + \|x_{k+1} - x_k\|^2 - \|y_k - x_{k+1}\|^2) + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\
& = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 + \frac{L}{2}\|y_k - x_{k+1}\|^2
\end{aligned} \tag{7}$$

Convergence rate for smooth and convex case

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

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

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

$$x^T y = \frac{1}{2}\left(\|x\|_2^2 + \|y\|_2^2 - \|x-y\|_2^2\right)$$

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 $x = x^*, y = y_k$:

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

$$\|y_k - x^*\|^2 \geq \|x^* - x_{k+1}\|^2 + \|y_k - x_{k+1}\|^2$$

$\pi_S(y_k) = x_{k+1}$

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:

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

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

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

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

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 lemma 7:

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease ~~lemma~~ 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 lemma 7:

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

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

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease lemma 7:

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

$$\le f(x_0)-f(x_k)+\cancel{\frac{L}{2}\sum_{i=0}^{i-1}\|y_i-x_{i+1}\|^2}+\frac{L}{2}\|x_0-x^*\|^2-\cancel{\frac{L}{2}\sum_{i=0}^{i-1}\|y_i-x_{i+1}\|^2}$$

$$\le f(x_0)-f(x_k)+\frac{L}{2}\|x_0-x^*\|^2$$

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease lemma 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 lemma 7:

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

$$\le f(x_0)-f(x_k)+\frac{L}{2}\sum_{i=0}^{i-1}\|y_i-x_{i+1}\|^2+\frac{L}{2}\|x_0-x^*\|^2-\frac{L}{2}\sum_{i=0}^{i-1}\|y_i-x_{i+1}\|^2$$

$$\le f(x_0)-f(x_k)+\frac{L}{2}\|x_0-x^*\|^2$$

$$\sum_{i=0}^{k-1}f(x_i)-kf^* \le f(x_0)-f(x_k)+\frac{L}{2}\|x_0-x^*\|^2$$

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

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

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

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 $x_{k+1} = \mathrm{proj}_S(y_k)$. By definition of projection,

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

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 $x_{k+1} = \mathrm{proj}_S(y_k)$. 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 $x_{k+1} = \mathrm{proj}_S(y_k)$. 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 $(*)$:

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

$-\frac{1}{2}$ $+\frac{1}{2}$

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 $x_{k+1} = \mathrm{proj}_S(y_k)$. 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

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

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

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

$k \cdot f_3 \leq f_1 + f_2 + f_3$

$k = 3$

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

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

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 $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex (or, PL-function with parameter μ) and L-smooth. Let $S \subseteq \mathbb{R}^n$ be a closed convex set, and assume that there is a minimizer x^* of f over S; The Projected Gradient Descent algorithm with stepsize $\frac{1}{L}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

$\left(\frac{L-\mu}{L}\right)$

1. We will use the notation $x_{k+1} = \text{proj}_S\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = \pi(y_k)$. Since projection is nonexpansive, we have:

$$\|x_{k+1} - x_k\| = \|\pi(y_k) - \pi(x_k)\| \leq \|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\| \quad \Rightarrow \quad \|x_{k+1} - x_k\|^2 \leq \frac{1}{L^2}\|\nabla f(x_k)\|^2 \qquad (8)$$

Convergence rate for smooth strongly convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex (or, PL-function with parameter μ) and L-smooth. Let $S \subseteq \mathbb{R}^n$ be a closed convex set, and assume that there is a minimizer x^* of f over S; The Projected Gradient Descent algorithm with stepsize $\frac{1}{L}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

1. We will use the notation $x_{k+1} = \text{proj}_S\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = \pi(y_k)$. Since projection is nonexpansive, we have:

$$\|x_{k+1} - x_k\| = \|\pi(y_k) - \pi(x_k)\| \leq \|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\| \quad \Rightarrow \quad \|x_{k+1} - x_k\|^2 \leq \frac{1}{L^2}\|\nabla f(x_k)\|^2 \tag{8}$$

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

(9)

Convergence rate for smooth strongly convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex (or, PL-function with parameter μ) and L-smooth. Let $S \subseteq \mathbb{R}^n$ be a closed convex set, and assume that there is a minimizer x^* of f over S; The Projected Gradient Descent algorithm with stepsize $\frac{1}{L}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

1. We will use the notation $x_{k+1} = \text{proj}_S\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = \pi(y_k)$. Since projection is nonexpansive, we have:

$$\|x_{k+1} - x_k\| = \|\pi(y_k) - \pi(x_k)\| \leq \|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\| \quad \Rightarrow \quad \|x_{k+1} - x_k\|^2 \leq \frac{1}{L^2}\|\nabla f(x_k)\|^2 \tag{8}$$

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

$$f(x_{k+1}) - f(x_k) \leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \tag{9}$$

Convergence rate for smooth strongly convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex (or, PL-function with parameter μ) and L-smooth. Let $S \subseteq \mathbb{R}^n$ be a closed convex set, and assume that there is a minimizer x^* of f over S; The Projected Gradient Descent algorithm with stepsize $\frac{1}{L}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

1. We will use the notation $x_{k+1} = \text{proj}_S\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = \pi(y_k)$. Since projection is nonexpansive, we have:

$$\|x_{k+1} - x_k\| = \|\pi(y_k) - \pi(x_k)\| \leq \|y_k - x_k\| = \frac{1}{L}\|\nabla f(x_k)\| \quad \Rightarrow \quad \|x_{k+1} - x_k\|^2 \leq \frac{1}{L^2}\|\nabla f(x_k)\|^2 \tag{8}$$

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

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{1}{2L}\|\nabla f(x_k)\|^2 \end{aligned} \tag{9}$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle \leq 0$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle \leq 0$$
$$\langle x_k - \tfrac{1}{L} \nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle \leq 0$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle \leq 0$$
$$\langle x_k - \tfrac{1}{L} \nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle \leq 0$$
$$\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L} \langle \nabla f(x_k), x_k - x_{k+1} \rangle \leq 0$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned} \langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\ \langle x_k - \tfrac{1}{L} \nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\ \langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L} \langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\ \langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L \|x_{k+1} - x_k\|^2 \end{aligned}$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned}
\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - \tfrac{1}{L} \nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L} \langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L \|x_{k+1} - x_k\|^2 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L \frac{1}{L^2} \|\nabla f(x_k)\|^2 = -\frac{1}{L} \|\nabla f(x_k)\|^2
\end{aligned}$$

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned}
\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - \tfrac{1}{L}\nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L}\langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\|x_{k+1} - x_k\|^2 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\frac{1}{L^2}\|\nabla f(x_k)\|^2 = -\frac{1}{L}\|\nabla f(x_k)\|^2
\end{aligned}$$

4. Returning back to the smoothness inequality (9) we obtain sufficient decrease property:

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned}
\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - \tfrac{1}{L}\nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L}\langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\|x_{k+1} - x_k\|^2 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\frac{1}{L^2}\|\nabla f(x_k)\|^2 = -\frac{1}{L}\|\nabla f(x_k)\|^2
\end{aligned}$$

4. Returning back to the smoothness inequality (9) we obtain sufficient decrease property:

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

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned}
\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - \tfrac{1}{L}\nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L}\langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\|x_{k+1} - x_k\|^2 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\frac{1}{L^2}\|\nabla f(x_k)\|^2 = -\frac{1}{L}\|\nabla f(x_k)\|^2
\end{aligned}$$

4. Returning back to the smoothness inequality (9) we obtain sufficient decrease property:

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

Convergence rate for smooth strongly convex case

3. By first-order optimality for that Euclidean projection problem, $\langle y_k - x_{k+1}, z - x_{k+1} \rangle \leq 0 \quad \forall z \in S$. In particular, for $z = x_k$,

$$\begin{aligned}
\langle y_k - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - \tfrac{1}{L}\nabla f(x_k) - x_{k+1}, x_k - x_{k+1} \rangle &\leq 0 \\
\langle x_k - x_{k+1}, x_k - x_{k+1} \rangle - \tfrac{1}{L}\langle \nabla f(x_k), x_k - x_{k+1} \rangle &\leq 0 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\|x_{k+1} - x_k\|^2 \\
\langle \nabla f(x_k), x_{k+1} - x_k \rangle &\leq -L\frac{1}{L^2}\|\nabla f(x_k)\|^2 = -\frac{1}{L}\|\nabla f(x_k)\|^2
\end{aligned}$$

4. Returning back to the smoothness inequality (9) we obtain sufficient decrease property:

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

Convergence rate for smooth strongly convex case

5. If we have a PL-function, we can use the following inequality: $\|\nabla f(x_k)\|^2 \geq 2\mu\,(f(x_k) - f^*)$. Thus, we can write:

$$f(x_{k+1}) \leq f(x_k) - \frac{\mu}{L}\,(f(x_k) - f^*)$$

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{\mu}{L}\,(f(x_k) - f^*)$$

Convergence rate for smooth strongly convex case

5. If we have a PL-function, we can use the following inequality: $\|\nabla f(x_k)\|^2 \geq 2\mu\,(f(x_k) - f^*)$. Thus, we can write:

$$f(x_{k+1}) \leq f(x_k) - \frac{\mu}{L}\,(f(x_k) - f^*)$$

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{\mu}{L}\,(f(x_k) - f^*)$$

6. Now we can use induction to get the following bound:

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k (f(x_0) - f^*)$$

Frank-Wolfe Method

Figure 11: Marguerite Straus Frank (1927-2024)

Figure 12: Philip Wolfe (1927-2016)

Idea

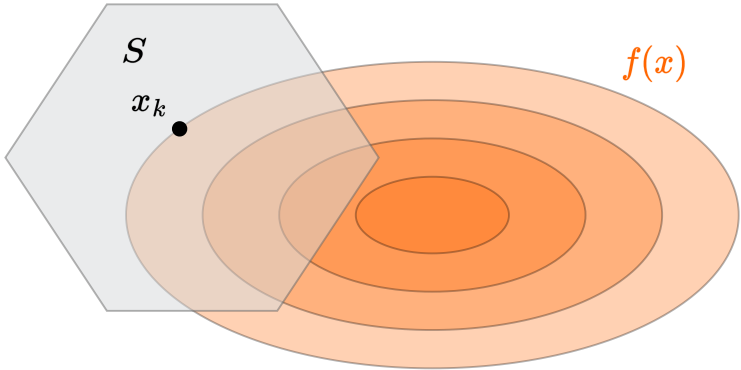


Figure 13: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

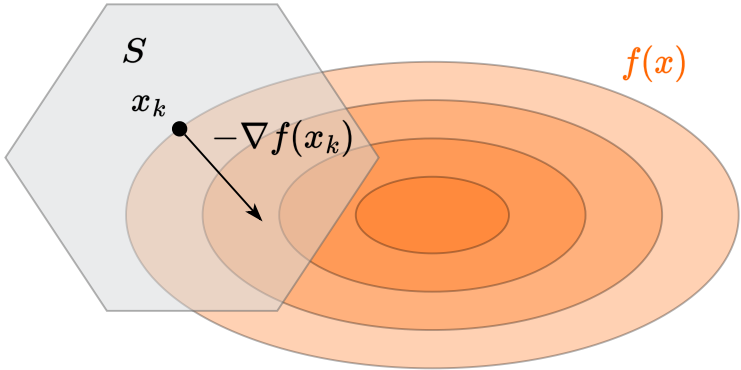


Figure 14: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

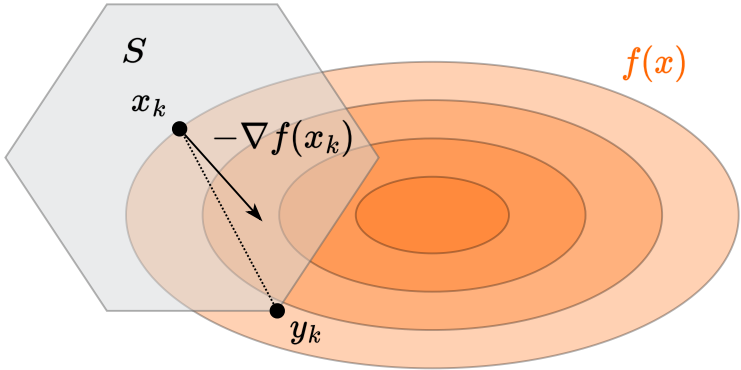


Figure 15: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

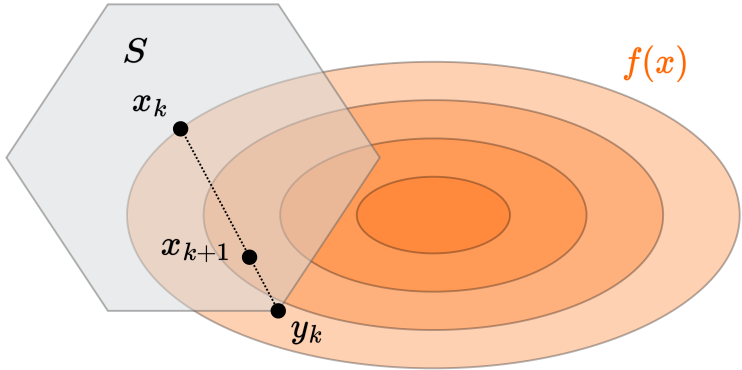


Figure 16: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

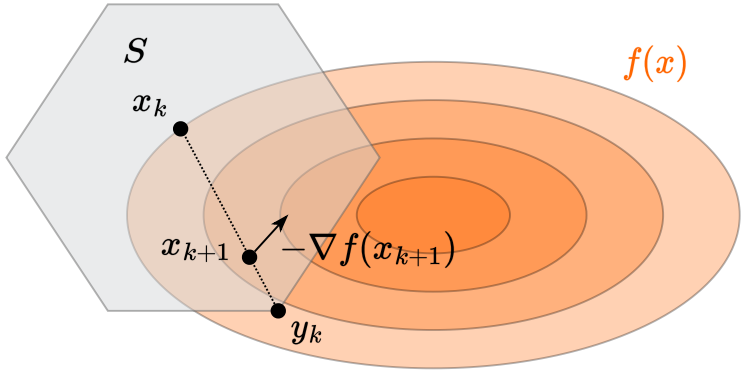


Figure 17: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

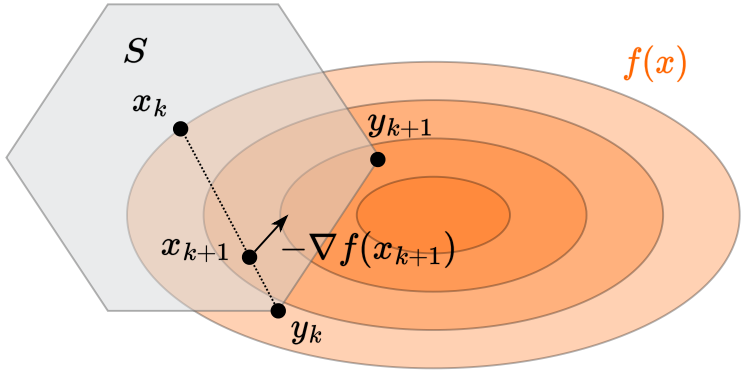


Figure 18: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

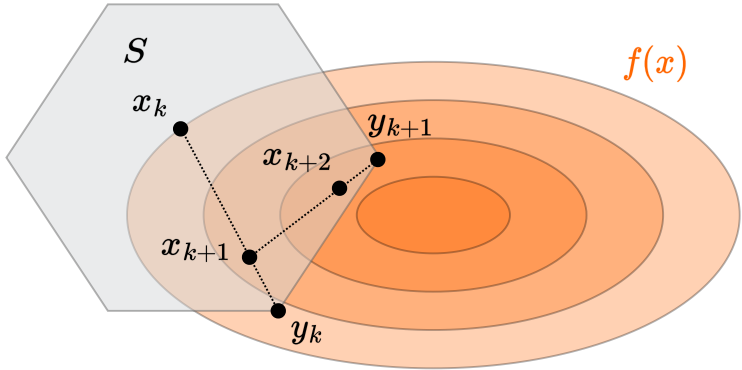


Figure 19: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea

$$f_{x_k}^I(x) = f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$$

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

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

LMO
Linear
Minimizat.
Oracle

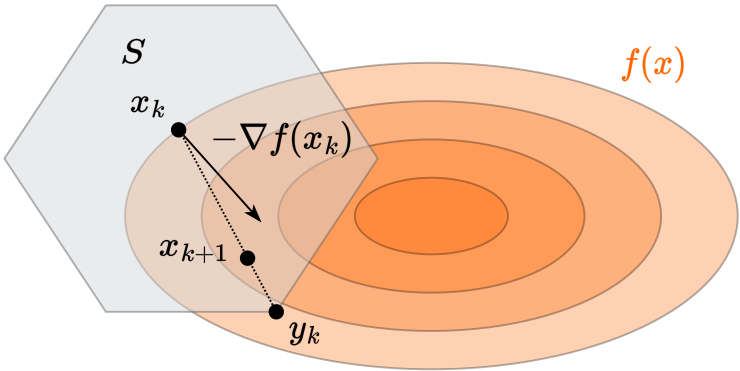


Figure 20: Illustration of Frank-Wolfe (conditional gradient) algorithm

Convergence rate for smooth and convex case

Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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:

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

В СИЛЬНО ВОГН. ТОЖЕ 1/K

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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ 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 $f(\hat{x}) - \min\limits_{x \in S} f(x) \leq \varepsilon$. The lower bound applies both for convex and strongly convex functions.

Sketch of the proof. Consider the following optimization problem:

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

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

Note, that:

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

[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

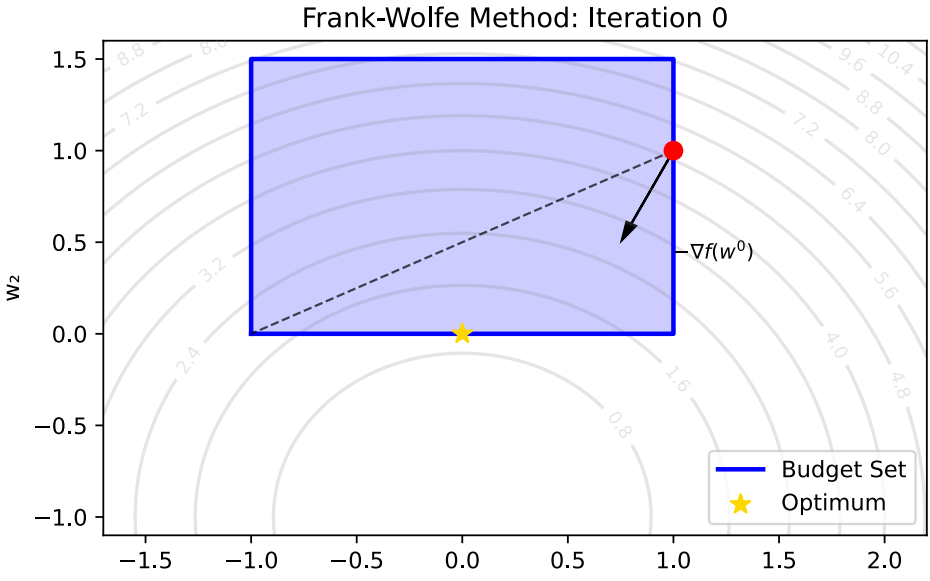
- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected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

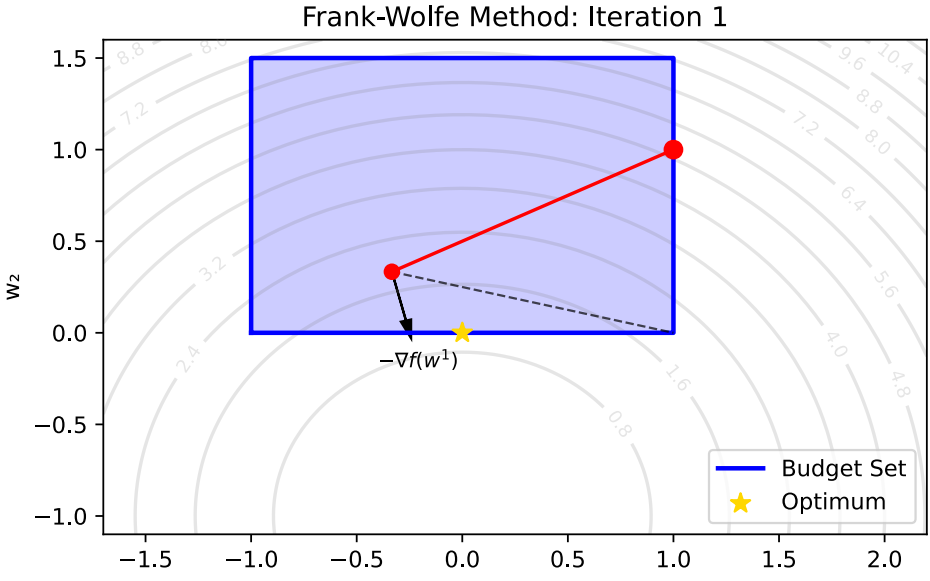
- Method does not require projections, in some special cases allows to compute iterations in closed form
- Global convergence rate is $O\left(\frac{1}{k}\right)$ for smooth and convex functions. Strong convexity does not improve the rate. This is the lower bound for LMO
- In comparison with projected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

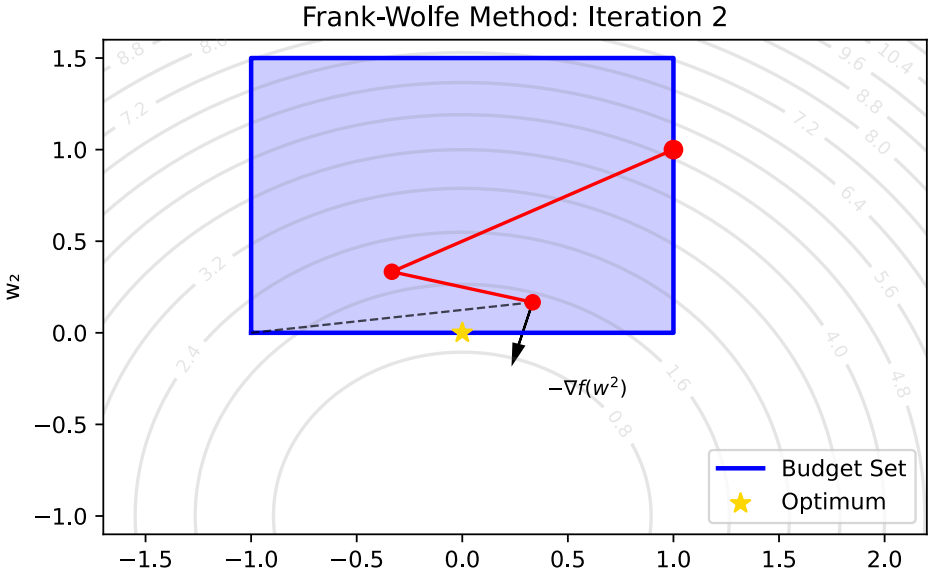
2d example. Frank-Wolfe method



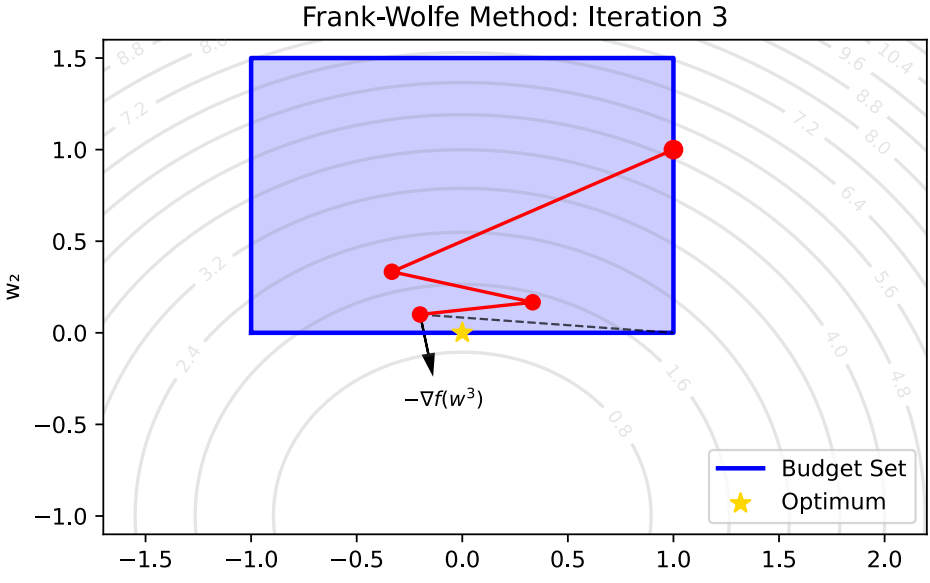
2d example. Frank-Wolfe method



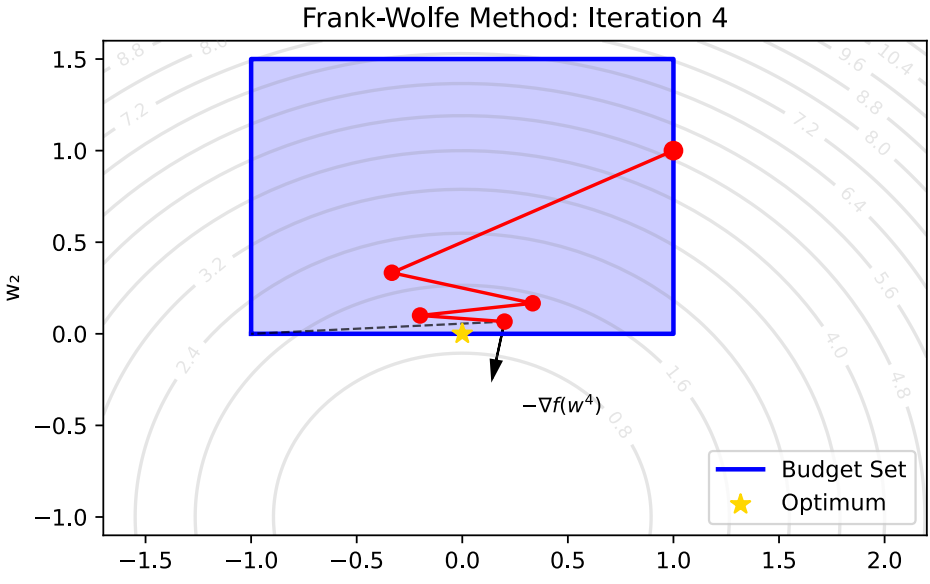
2d example. Frank-Wolfe method



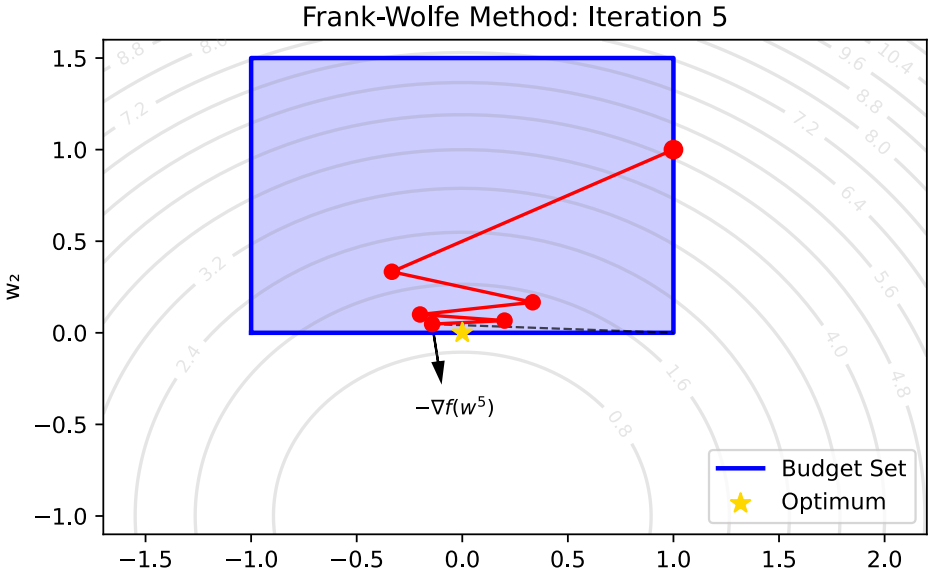
2d example. Frank-Wolfe method



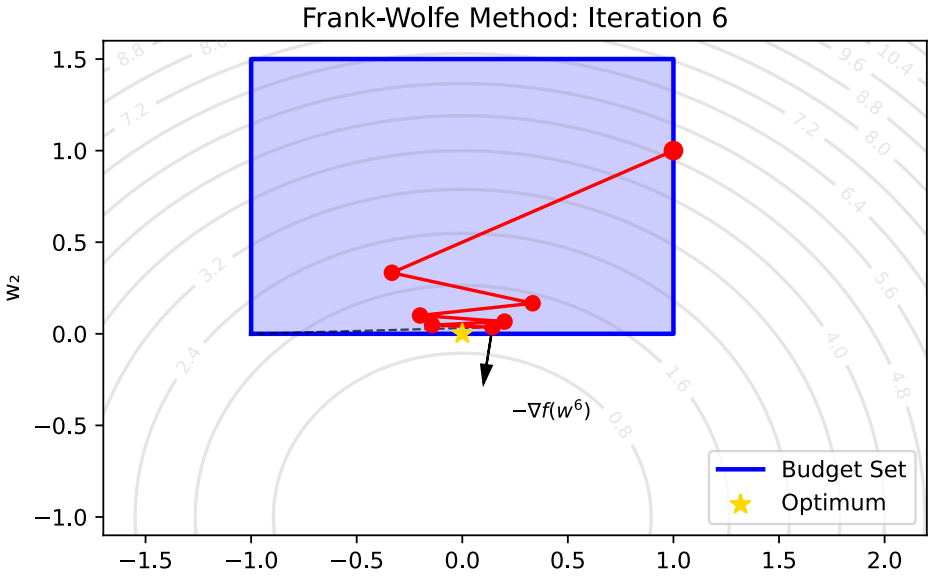
2d example. Frank-Wolfe method



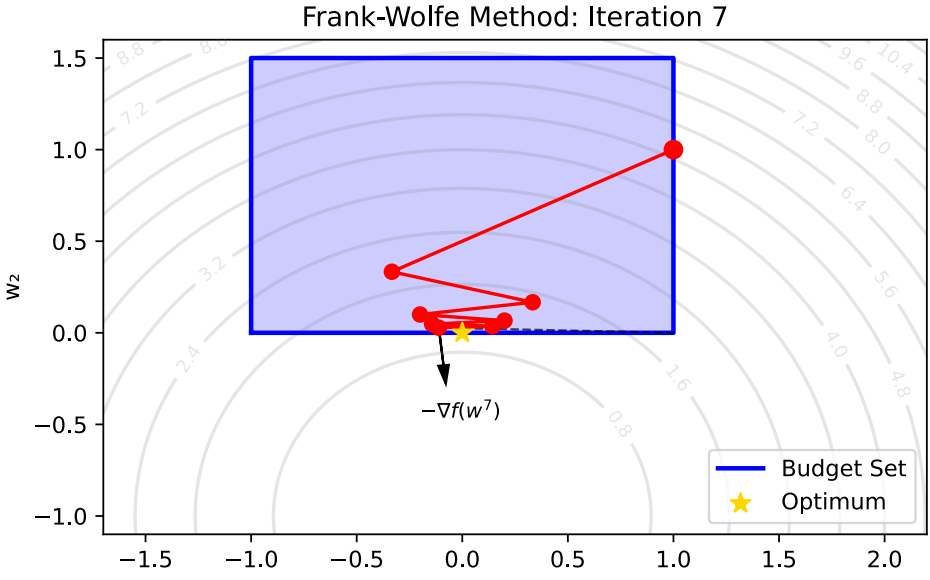
2d example. Frank-Wolfe method



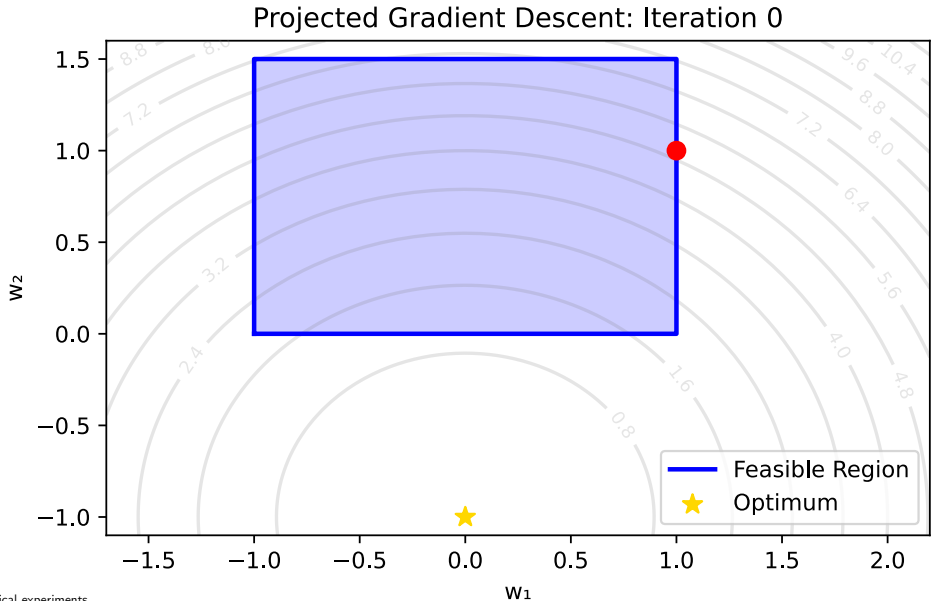
2d example. Frank-Wolfe method



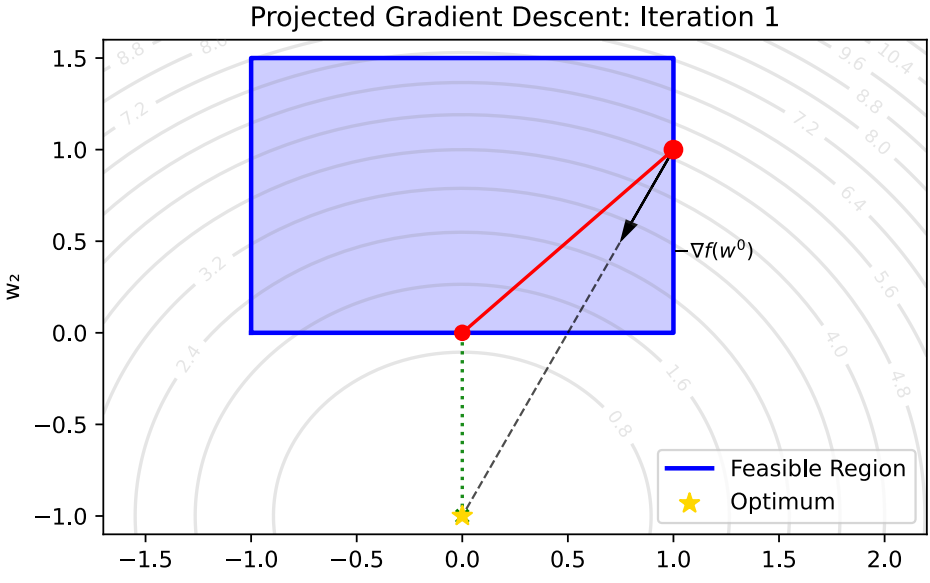
2d example. Frank-Wolfe method



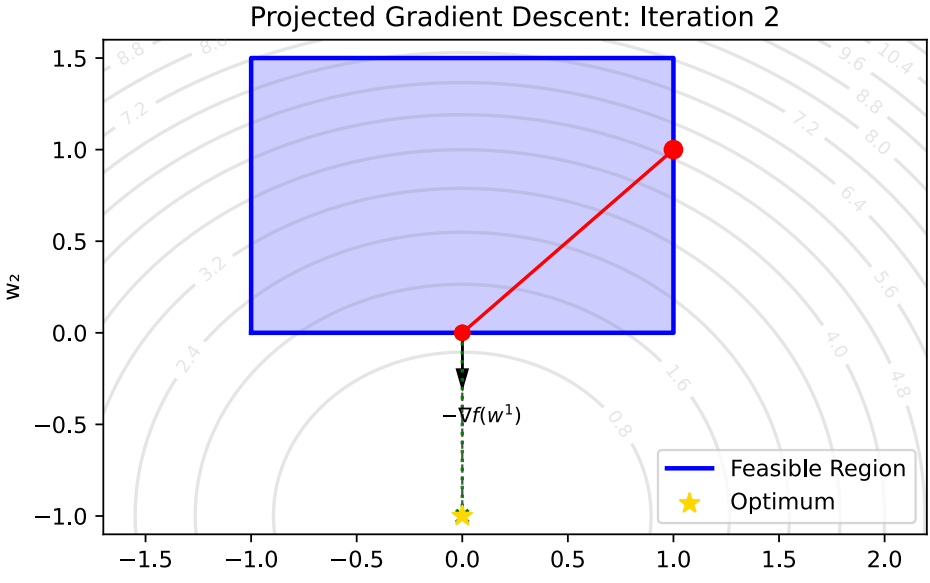
2d example. Projected gradient descent



2d example. Projected gradient descent



2d example. Projected gradient descent



Quadratic function. Box constraints

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

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

The projection is simple:

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

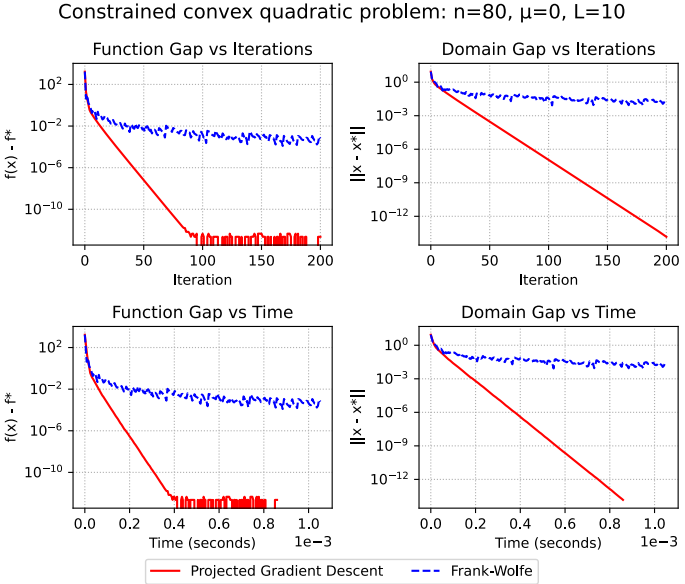
or

$$\pi_S(x) = \max\left(-\mathbf{1}, \min(\mathbf{1}, x)\right).$$

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

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

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



Quadratic function. Box constraints

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

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

The projection is simple:

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

or

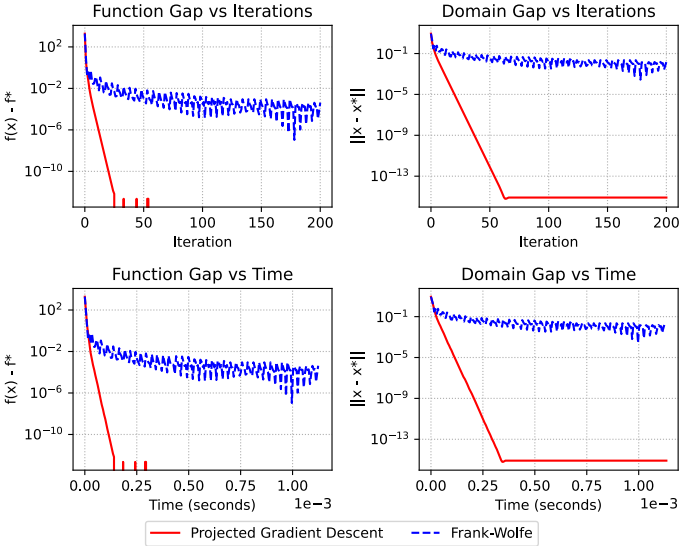
$$\pi_S(x) = \max\left(-\mathbf{1}, \min(\mathbf{1}, x)\right).$$

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

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, μ=1, L=10



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

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

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

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

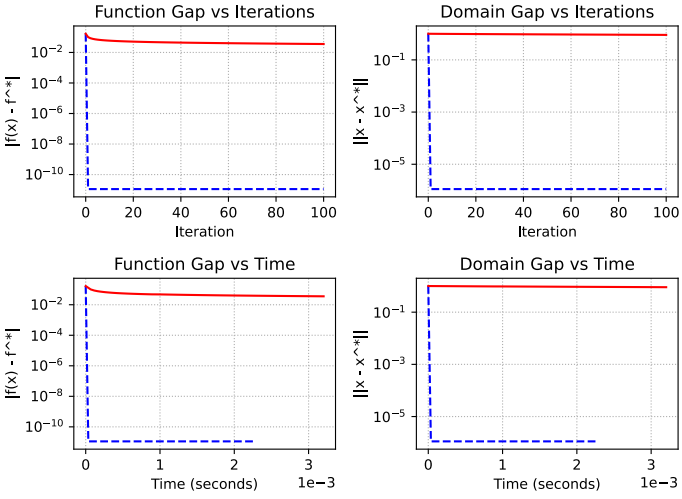
The projection onto the unit simplex $\pi_S(x)$ can be done in $\mathcal{O}(n \log n)$ or expected $\mathcal{O}(n)$ time. [a]

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:

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

[a] Efficient Projections onto the ℓ_1-Ball for Learning in High Dimensions

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



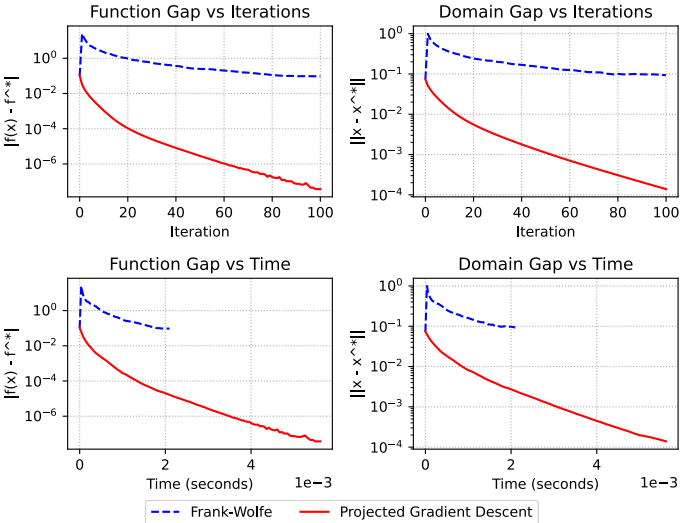
Quadratic function. Simplex constraints

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

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

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

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



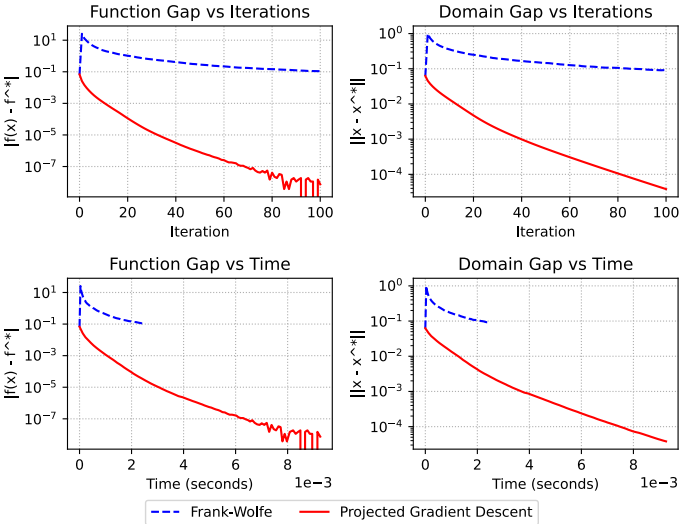
Quadratic function. Simplex constraints

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

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

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

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



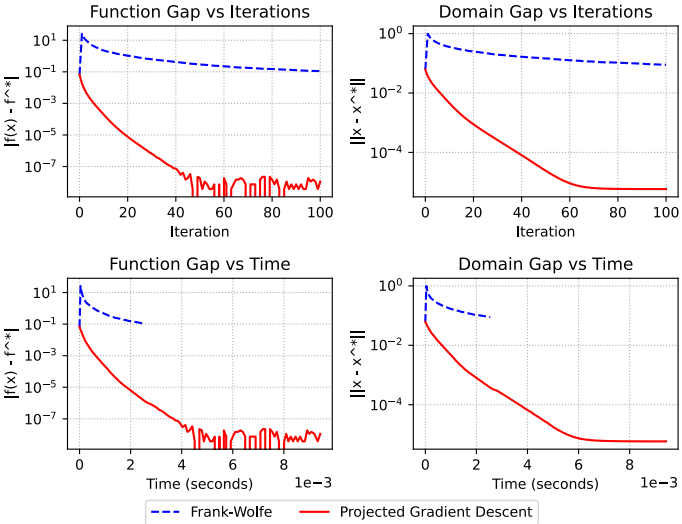
Quadratic function. Simplex constraints

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

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

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

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



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

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 thedistancetothe optimum. For the nuclear norm ball, i.e., the spectrahedron, ν denotes the number of non-zero entries and σ_1 denotes the top singular value of the projected matrix.