

A Corgi dog and a yellow rubber duck are positioned inside a transparent, wireframe cube. The Corgi is on the left, looking towards the camera, and the rubber duck is on the right, also looking towards the camera. The cube is made of thin, metallic-looking lines. The background is a plain, light-colored surface.

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

Daniil Merkulov

Optimization methods. MIPT

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



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



- 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



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



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

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



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

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

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

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

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

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

- **Sufficient conditions of existence of a projection.** If $S \subseteq \mathbb{R}^n$ - closed set, then the projection on set S exists for any point.
- **Sufficient conditions of uniqueness of a projection.** If $S \subseteq \mathbb{R}^n$ - closed convex set, then the projection on set S is unique for any point.
- If a set is open, and a point is beyond this set, then its projection on this set 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

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

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

Proof

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)

i Theorem

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

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

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

Proof

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)

i Theorem

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

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

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

Proof

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.

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

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



Figure 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

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

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

Proof

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.

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

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

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



Figure 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

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

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

Proof

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:



Figure 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

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

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

Proof

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:

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



Figure 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

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

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

Proof

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:

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

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



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

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

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 Equation 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 Equation 3

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

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

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

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

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

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

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

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 Equation 3

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

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

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

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

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

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

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

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

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

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 Equation 3

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

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

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

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

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

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

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

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

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

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

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

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:

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

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:

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

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



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:



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

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



Figure 10: Illustration of Projected Gradient Descent algorithm

Convergence rate for smooth and convex case

i 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

i 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 $k > 0$:

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

Proof

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

i 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 $k > 0$:

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

Proof

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

i 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 $k > 0$:

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

Proof

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

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

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

(7)

Convergence rate for smooth and convex case

i 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 $k > 0$:

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

Proof

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

i 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 $k > 0$:

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

Proof

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

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

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

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

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

Convergence rate for smooth and convex case

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

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

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

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

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:

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

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

Convergence rate for smooth and convex case

5. Bound gradients with sufficient decrease lemma 7:

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

Convergence rate for smooth and convex case

6. From the sufficient decrease inequality

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

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

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

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.

Frank-Wolfe Method

Idea



Figure 11: Illustration of Frank-Wolfe (conditional gradient) algorithm

Idea



Figure 12: Illustration of Frank-Wolfe (conditional gradient) algorithm

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

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

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



Figure 18: 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

i 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 .

Proof

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. Starting from $\delta_2 \leq \frac{1}{2}$ and applying induction on k , we can show that:

$$\delta_k \leq \frac{2}{k+1}$$

which gives us the desired result:

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

Convergence rate for strongly convex case

i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly 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 $\gamma_k = \frac{2}{k+2}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \frac{4LR^2}{(k+2)^2}$$

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

Convergence rate for strongly convex case

i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly 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 $\gamma_k = \frac{2}{k+2}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \frac{4LR^2}{(k+2)^2}$$

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

Proof

1. By μ -strong convexity of f , for any $x, y \in S$:

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

Convergence rate for strongly convex case

i Theorem

Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be μ -strongly 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 $\gamma_k = \frac{2}{k+2}$ achieves the following convergence after iteration $k > 0$:

$$f(x_k) - f^* \leq \frac{4LR^2}{(k+2)^2}$$

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

Proof

1. By μ -strong convexity of f , for any $x, y \in S$:

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

2. This gives us a stronger inequality than in the convex case:

$$\langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*) - f(x_k) - \frac{\mu}{2} \|x^* - x_k\|^2$$

Convergence rate for strongly convex case

3. Following similar steps as in the convex case, but using the stronger inequality:

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

Convergence rate for strongly convex case

3. Following similar steps as in the convex case, but using the stronger inequality:

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

4. Rearranging terms and using the fact that $\|x^* - x_k\|^2 \geq 0$:

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2} - (1 - \gamma_k) \frac{\mu}{2} \|x^* - x_k\|^2 \\ &\leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2} \end{aligned}$$

Convergence rate for strongly convex case

3. Following similar steps as in the convex case, but using the stronger inequality:

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

4. Rearranging terms and using the fact that $\|x^* - x_k\|^2 \geq 0$:

$$\begin{aligned} f(x_{k+1}) - f(x^*) &\leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2} - (1 - \gamma_k) \frac{\mu}{2} \|x^* - x_k\|^2 \\ &\leq \gamma_k (f(x_k) - f(x^*)) + (1 - \gamma_k)^2 \frac{LR^2}{2} \end{aligned}$$

5. With $\gamma_k = \frac{2}{k+2}$ and denoting $\delta_k = f(x_k) - f^*$, we get:

$$\begin{aligned} \delta_{k+1} &\leq \frac{2}{k+2} \delta_k + \frac{LR^2}{2} \left(1 - \frac{2}{k+2} \right)^2 \\ &= \frac{2}{k+2} \delta_k + \frac{LR^2}{2} \frac{(k)^2}{(k+2)^2} \end{aligned}$$

Convergence rate for strongly convex case

6. It can be shown by induction that:

$$\delta_k \leq \frac{4LR^2}{(k+2)^2}$$

Convergence rate for strongly convex case

6. It can be shown by induction that:

$$\delta_k \leq \frac{4LR^2}{(k+2)^2}$$

7. This gives us the improved convergence rate of $\mathcal{O}(\frac{1}{k^2})$ for the strongly convex case, compared to $\mathcal{O}(\frac{1}{k})$ for the convex case.