Numerical Optimization

Lecture 7: Gradient Projection Method, Conditional Gradient Method

Hao Wang
Email: wanghao1@shanghaitech.edu.cn

Contents

We have learned Gradient Descent Method. But there are more...

- Subgradient descent method
- Proximal gradient method
- Coordinate descent method
- Projected gradient, projected subgradient method
- Frank-Wolfe method (Conditional gradient method)
- Nonlinear Conjugate Gradient, CG
- Stochastic gradient descent method (SGD), Momentum SGD, Adam (Adaptive Moment Estimation), AdaGrad (Adaptive Gradient), RMSProp
- Nesterov acceleration, extrapolation/interapolation acceleration
- MUON

Contents

Recall, the effective domain of f is $\text{dom}(f) := \{x \mid x \in \mathcal{X} \text{ and } f(x) < \infty\}$.

Definition 1 (Subgradient and Subdifferential)

A vector $g \in \mathbb{R}^n$ is a subgradient of a proper convex f at $x \in \text{dom}(f)$ if

$$f(\bar{x}) \geq f(x) + g^T(\bar{x} - x) \text{ for all } \bar{x} \in \mathbb{R}^n.$$

The set of all subgradients of f at x, denoted $\partial f(x)$, is the subdifferential of f at x.

Theorem 2

If f is convex, then x^* is the global minimizer $\iff$ $0 \in \partial f(x^*)$.

A function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is proper if

$$f(x) \begin{cases} < +\infty & \text{for some } x \in \mathcal{X} \\ > -\infty & \text{for all } x \in \mathcal{X}. \end{cases}$$

Otherwise, it is improper.

Let $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ be proper and convex.

- If $x \in \text{dom}(f)$, then $g \in \partial f(x)$ if and only if [1]

$$f'(d; x) \geq g^T d \text{ for all } d \in \mathbb{R}^n.$$

- If $x \in \text{int dom}(f)$, then $\partial f(x)$ is a nonempty, convex, and compact and

$$f'(d; x) = \max_{g \in \partial f(x)} g^T d \text{ for all } d \in \mathbb{R}^n.$$

[1] Proof: $f(x + td) - f(x) \geq tg^T d$

Theorem 3

x^* is the minimizer of $f \implies f'(d; x^*) \geq 0, \forall d$. If f is convex, it's $\iff$.

- At a point $x \in \mathbb{R}^n$, a descent direction d is one for which we have

$$\sup_{g \in \partial f(x)} g^T d = f'(d; x) < 0.$$

- For proper convex f and $x \in \text{int } (\text{dom}(f))$, we find that

$$\begin{aligned} \min_{\|d\|_2 \leq 1} f'(d; x) &= \min_{\|d\|_2 \leq 1} \max_{g \in \partial f(x)} g^T d \\ &= \max_{g \in \partial f(x)} \min_{\|d\|_2 \leq 1} = \max_{g \in \partial f(x)} (-\|g\|_2) = - \min_{g \in \partial f(x)} \|g\|_2, \end{aligned}$$

that is, in such a case, the steepest descent direction is $d = -g/\|g\|_2$, where g is the minimum norm element in $\partial f(x)$

$$\min_x \psi(x) = f(x) + h(x)$$

where $f \in \mathcal{C}$ but not necessarily convex, and h is convex.

Theorem 4

If x^* is a local minimizer, then $-\nabla f(x^*) \in \partial h(x^*)$.

Proof.

Suppose by contradiction $-\nabla f(x^*) \notin \partial h(x^*)$. By the Separation Theorem, $\exists d$ and scalar b such that

$$\theta^T d + b < 0 < -\nabla f(x^*)^T d + b, \ \forall \theta \in \partial h(x^*).$$

By the definition of directional derivative

$$\psi'(x^*; d) \leq \nabla f(x^*)^\top d + \sup_{\theta \in \partial h(x^*)} \theta^\top d < 0.$$

□

We often want to optimize f within a feasible set Ω:

$$\min f(x), \quad \text{s.t. } x \in \Omega.$$

- $x^* \in \Omega$ is a global solution if

$$f(x^*) \leq f(x) \quad \forall x \in \Omega.$$

- $x^* \in \Omega$ is a local solution if there is a neighborhood $\mathcal{N}$ of x^* s.t.

$$f(x^*) \leq f(x) \quad \forall x \in \mathcal{N} \cap \Omega.$$

- $x^* \in \Omega$ is a strict/strong local solution if there is a neighborhood $\mathcal{N}$ of x^* s.t.

$$f(x^*) < f(x) \quad \forall x \in (\mathcal{N} \cap \Omega) \setminus x^*.$$

Consider a general constrained optimization over a closed convex set Ω:

$$\min f(x), \quad \text{s.t. } x \in \Omega.$$

Write it equivalently as

$$\min_x f(x) + I_\Omega(x),$$

where I_Ω is the indicator function

$$I_\Omega(x) = \begin{cases} 0 & \text{if } x \in \Omega \\ +\infty & \text{if } x \notin \Omega \end{cases}$$

Notably, I_Ω is a convex function!
If x^* is the minimizer, then

$$-\nabla f(x^*) \in \partial I_\Omega(x^*).$$

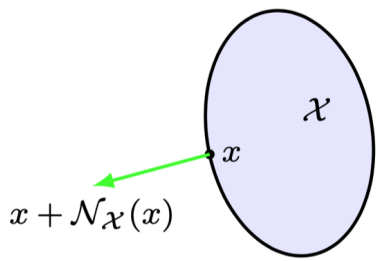
But, what is this $\partial I_\Omega(x)$?

Theorem 5 (Normal Cone)

Given a nonempty convex $\Omega \subset \mathbb{R}^n$ and $x \in \Omega$, the normal cone of Ω at x is

$$\mathcal{N}_\Omega(x) := \{g \mid g^T(\bar{x} - x) \leq 0 \text{ for all } \bar{x} \in \Omega\}.$$

If $x \in \text{int}(\Omega)$, then clearly $\mathcal{N}_\Omega(x) = \{0\}$, but for $x \notin \text{int}(\Omega)$ it contains at least one halfline.



$$I_\Omega(x) \geq I_\Omega(x^*) + g^T(x - x^*) \iff g \in \mathcal{N}_\Omega(x^*).$$

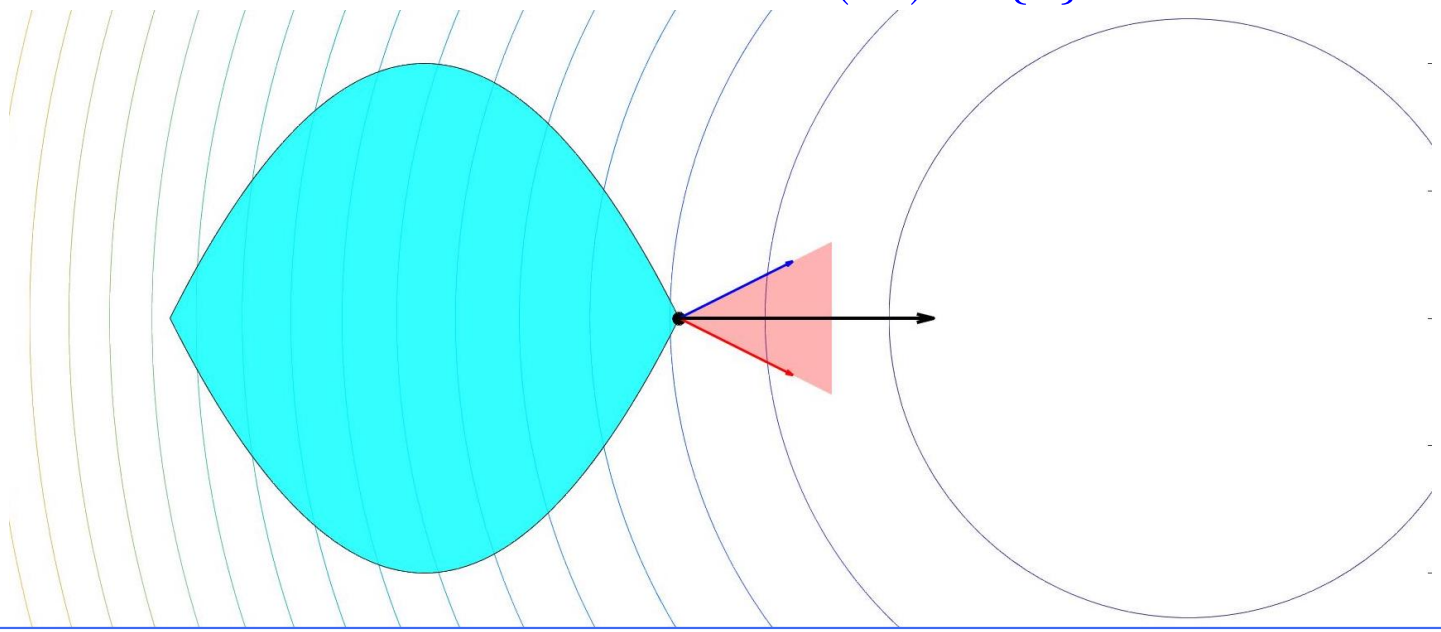
$$\partial I_\Omega(x^*) = \mathcal{N}_\Omega(x^*).$$

Theorem 6

If x^* is a minimizer of $f \in \mathcal{C}$ in Ω, then

$$-\nabla f(x^*) \in \mathcal{N}_\Omega(x^*).$$

That is, if x^* is a minimizer, then the steepest descent direction for f at x^* is in the normal cone of Ω at x^*. (Blue vector denotes $\nabla f(x^*)$.)
In the unconstrained case, $\Omega = \mathbb{R}^n$ and $\mathcal{N}_\Omega(x^*) = \{0\}$.



Proof.

Suppose x^* is a local minimizer. It follows that $f(z) \geq f(x^*)$ with $z = x^* + \tau(y - x^*) \in \Omega$ for any $y \in \Omega$ and sufficiently small $\tau \in (0, 1)$. Therefore, there exists $\tau' \in (0, \tau)$

$$f(z) = f(x^*) + \nabla f(x^* + \tau'(y - x^*))^T \tau(y - x^*) \geq f(x^*),$$

implying

$$\nabla f(x^* + \tau'(y - x^*))^T (y - x^*) \geq 0.$$

Letting $\tau \to 0$, since $f \in \mathcal{C}$,

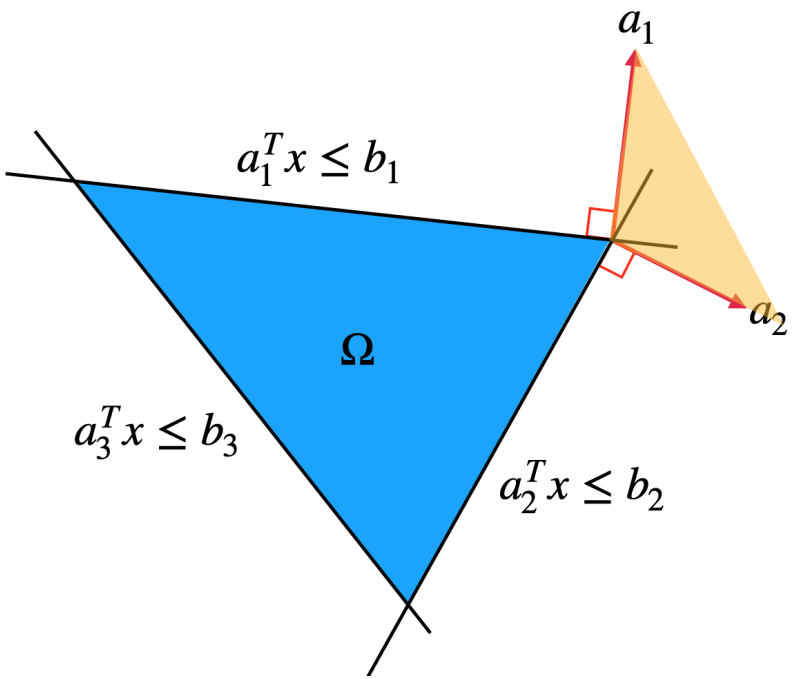
$$\nabla f(x^*)^T (y - x^*) \geq 0.$$

By the definition of normal cone,

$$-\nabla f(x^*) \in \mathcal{N}_\Omega(x^*).$$

- Unconstrained case $\Omega = \mathbb{R}^n \implies N_\Omega(x^*) = \{0\}$.
  Optimality condition: $-\nabla f(x^*) = 0 \quad \Rightarrow \quad \nabla f(x^*) = 0.$

- Linear Inequality Constraints: $\Omega = \{x \mid Ax \leq b\}$



$\mathcal{N}_\Omega(x^*) = \{\lambda_1 a_1 + \lambda_2 a_2 \mid \lambda_1 \geq 0, \lambda_2 \geq 0\}$

- Linear Inequalities: $\Omega = \{x \mid Ax \leq b,\ A \in \mathbb{R}^{m \times n}\}$
- Generally, $\mathcal{N}_\Omega(x^*) = \{\sum_{i=1}^{m} \lambda_i a_i \mid \lambda_i \geq 0, \lambda_i(a_i^T x^* - b_i) = 0\}$
- $$-\nabla f(x^*) \in \mathcal{N}_\Omega(x^*) \implies \begin{cases} \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i a_i = 0, \\ a_i^T x^* \leq b_i, \\ \lambda_i \geq 0,\ \lambda_i(a_i^T x^* - b_i) = 0 \end{cases}$$

  Familiar? This is called the KKT (Karush-Kuhn-Tucker) conditions!

- With equalities $\Omega = \{x \mid Ax \leq b, Cx = d,\ A \in \mathbb{R}^{m \times n}, C \in \mathbb{R}^{l \times n}\}$

$$-\nabla f(x^*) \in \mathcal{N}_\Omega(x^*) \implies \begin{cases} \nabla f(x^*) + \sum_i \lambda_i a_i + \sum_j \mu_j c_j = 0, \\ a_i^T x^* \leq b_i,\ c_i^T x^* = d_i \\ \lambda_i \geq 0,\ \lambda_i(a_i^T x^* - b_i) = 0 \end{cases}$$

Contents

Suppose $\min_x \ f(x)$

Construct local approximation at x^k:

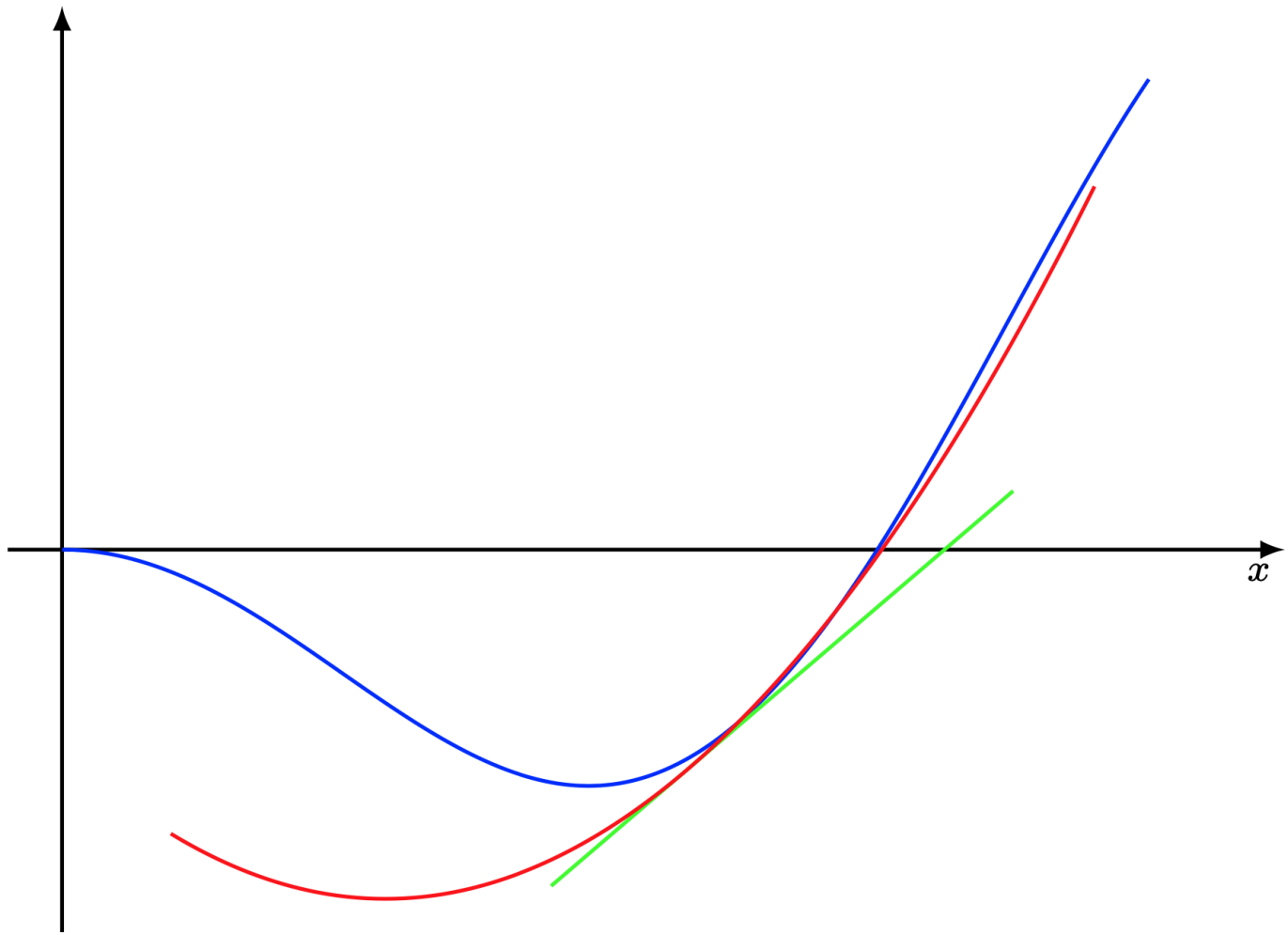
$$x^{k+1} \leftarrow \min_x \ f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2t}\|x - x^k\|_2^2$$

$$\implies x^{k+1} \leftarrow x^k - t\nabla f(x^k)$$

Interpretation:

- Go along the steepest descent direction with a stepsize t
- Linear model but not too far away from x^k
- Local quadratic approximation model at x^k
- Newton method, quasi-Newton method also use local models.

$$x^{k+1} \leftarrow \min_x \; f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2t} \|x - x^k\|_2^2$$



$$\min f(x) \quad \text{s.t. } x \in \Omega$$

where Ω is closed convex set.

Construct local approximation at x^k:

$$\begin{cases} \min & f(x^k) + \nabla f(x^k)^T (x - x^k) + \frac{1}{2t} \|x - x^k\|_2^2 \\ \text{s.t.} & x \in \Omega. \end{cases}$$

Interpretation:

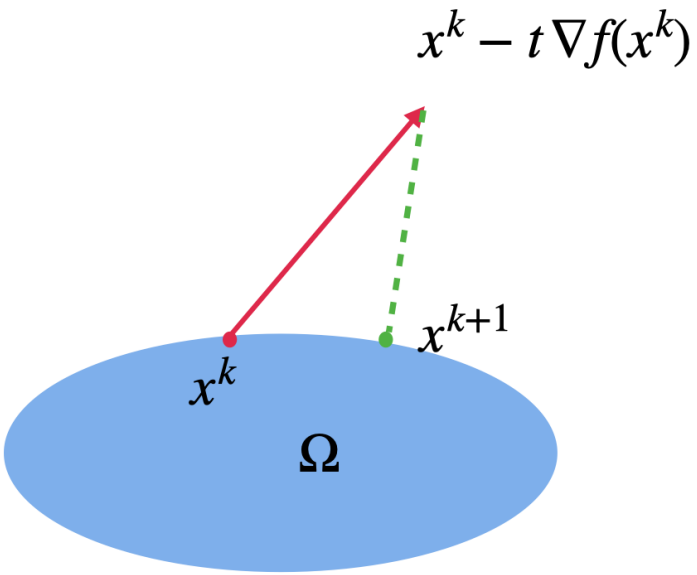
- Go along the steepest descent direction with a stepsize t but don't go outside
- Local quadratic approximation model at x^k

Construct local approximation at x^k:

$$\begin{cases} \min \left\| x - [x^k - t\nabla f(x^k)] \right\|_2^2 \\ \text{s.t.} \quad x \in \Omega. \end{cases}$$

Interpretation:

- Find the point in Ω that is the closest to the GD iterate.
- This is called a projection.



Theorem 7

Let $x^0 \in \mathbb{R}^n$ and let $\Omega \subset \mathbb{R}^n$ be a nonempty closed convex set. Then $\bar{x} \in \Omega$ solves the problem

$$\min_x \tfrac{1}{2}\|x - x^0\|_2^2 \quad \textit{s.t. } x \in \Omega$$

if and only if $(\bar{x} - x_0)^T(y - \bar{x}) \geq 0$ for all $y \in \Omega$. Moreover, the solution $\bar{x}$ always exists and is unique.

Proof.

Existence follows from the compactness of the set

$$\{x \in \Omega : \|x - x^0\|_2 \leq \|\hat{x} - x^0\|_2\}$$

where $\hat{x}$ is any element of C. Uniqueness follows from the strong convexity of the 2-norm squared. From the optimality condition $-(\bar{x} - x^0) \in \mathcal{N}_\Omega(x^0)$, meaning

$$(\bar{x} - x^0)^T(y - \bar{x}) \geq 0$$

□

Proposition 8 (Nonexpansiveness)

$$\|P_\Omega(x) - P_\Omega(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n$$

Proposition 9 (Variational Inequality)

For all $y \in \Omega$ *and all* $z \in \mathbb{R}^n$*:*

$$(z - P_\Omega(z))^\top (P_\Omega(z) - y) \geq 0$$

Proposition 10 (Fixed Point Characterization)

x^* *is optimal* $\iff$ $x^* = P_\Omega(x^* - \alpha \nabla f(x^*))$ *for any* $\alpha > 0$

Examples:

- $\Omega = \{x \mid l_i \leq x_i \leq u_i\}$
- $\Omega = \{x \mid \|x\|_2 \leq R\}$
- $\Omega = \{x \mid x_i \geq 0\}$
- $\Omega = \{x \mid a^T x = b\}$
- $\Omega = \{x \mid a^T x \leq b\}$

$$\min_x \quad f(x), \quad \text{s.t. } x \in \Omega.$$

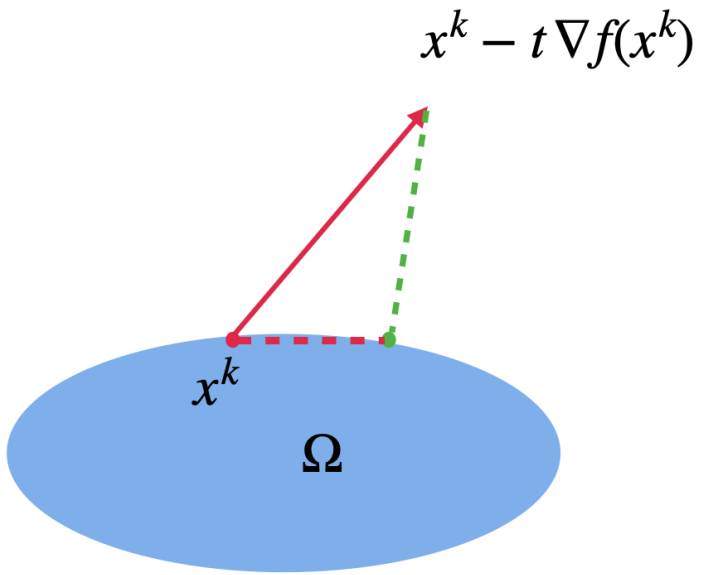
f is $\mathcal{C}^1$, C is closed convex

Gradient Projection Algorithm:

Set $d^k = P_\Omega(x^k - \nabla f(x^k)) - x^k$

Set λ^k by backtracking Armijo line search

Set $x^{k+1} \leftarrow x^k + \lambda^k d^k$



Proposition 11

Let $x \in \Omega$ and set $d = P_\Omega(x - t\nabla f(x)) - x$. Then

$$\nabla f(x)^T d \leq -\frac{\|P_\Omega(x - t\nabla f(x)) - x\|^2}{t}.$$

Proof.

Let $z = P_\Omega(x - t\nabla f(x)), d = z - x$. Simply observe that

$$\begin{aligned}\|P_\Omega(x - t\nabla f(x)) - x\|^2 &= \langle z - x, z - x\rangle \\ &= -t\nabla f(x)^T d + \langle z - (x - t\nabla f(x)), z - x\rangle \\ &\leq -t\nabla f(x)^T d.\end{aligned}$$

□

Apply the Zoutendijk's result to have

$$\frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \to 0$$

Therefore, combining the above Proposition to yield

$$\frac{\|P_\Omega(x - t\nabla f(x)) - x\|^4}{t^2\|d^k\|^2} = \|d^k\|^2/t^2 \to 0$$

Therefore, $P_\Omega(x^k - t\nabla f(x^k)) - x^k \to 0$.

Every limit point satisfies $P_\Omega(x - t\nabla f(x)) - x = 0$

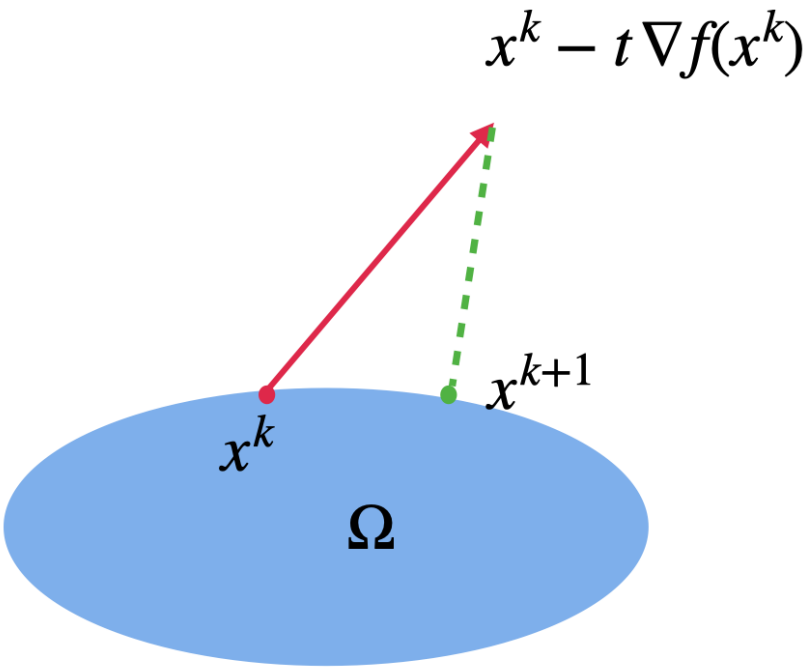
$$x - t\nabla f(x) - x \in \mathcal{N}_\Omega(x) \implies -\nabla f(x) \in \mathcal{N}_\Omega(x)$$

Consider the constrained optimization problem:

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

Gradient Projection Method (fixed step size $\alpha > 0$):

$$x^{k+1} = P_\Omega\left(x^k - \alpha \nabla f(x^k)\right)$$



Theorem 12 (Global Convergence of Gradient Projection)

Assume:

1. f *has* L*-Lipschitz gradient on* Ω*:* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$
2. f *is* μ*-strongly convex on* Ω $(\mu > 0)$
3. *Step size satisfies* $0 < \alpha < \frac{2}{L}$

Then the gradient projection method with fixed step size α *converges globally to the unique minimizer* x^* *with linear rate:*

$$\|x^{k+1} - x^*\| \leq \sqrt{q(\alpha)}\|x^k - x^*\|$$

where $q(\alpha) = 1 - 2\alpha\mu + \alpha^2 L^2 < 1$ *for* $\alpha \in (0, \frac{2\mu}{L^2})$.

Proof.

Let x^* be the unique minimizer. Define:

$$z^k = x^k - \alpha \nabla f(x^k), \quad z^* = x^* - \alpha \nabla f(x^*)$$

Using the projection's nonexpansiveness:

$$\|x^{k+1} - x^*\|^2 = \|P_\Omega(z^k) - P_\Omega(z^*)\|^2 \leq \|z^k - z^*\|^2$$

Expanding the right-hand side:

$$\|z^k - z^*\|^2 = \|(x^k - x^*) - \alpha(\nabla f(x^k) - \nabla f(x^*))\|^2$$

□

Proof.

$$\begin{aligned}
&\|z^k - z^*\|^2 \\
&= \|x^k - x^*\|^2 - 2\alpha(x^k - x^*)^\top(\nabla f(x^k) - \nabla f(x^*)) \\
&\quad + \alpha^2\|\nabla f(x^k) - \nabla f(x^*)\|^2.
\end{aligned}$$

Using strong convexity:

$$(x^k - x^*)^\top(\nabla f(x^k) - \nabla f(x^*)) \geq \mu\|x^k - x^*\|^2$$

Using Lipschitz gradient:

$$\|\nabla f(x^k) - \nabla f(x^*)\|^2 \leq L^2\|x^k - x^*\|^2$$

□

Proof.

Substituting these inequalities:

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\alpha\mu\|x^k - x^*\|^2 + \alpha^2 L^2 \|x^k - x^*\|^2$$

Thus:

$$\|x^{k+1} - x^*\|^2 \leq \left[1 - 2\alpha\mu + \alpha^2 L^2\right] \|x^k - x^*\|^2$$

Let $q(\alpha) = 1 - 2\alpha\mu + \alpha^2 L^2$. For $\alpha \in (0, \frac{2\mu}{L^2})$, we have:

$$q(\alpha) < 1 \quad \text{and} \quad \|x^{k+1} - x^*\| \leq \sqrt{q(\alpha)}\|x^k - x^*\|$$

This establishes global linear convergence to the unique minimizer x^*. □

Optimal Step Size

The contraction factor $q(\alpha) = 1 - 2\alpha\mu + \alpha^2 L^2$ is minimized at:

$$\alpha^* = \frac{\mu}{L^2}$$

with optimal value:

$$q^* = 1 - \frac{\mu^2}{L^2} = 1 - \frac{1}{\kappa^2}, \quad \kappa = \frac{L}{\mu} \text{ (condition number)}$$

Linear Convergence Rate

With optimal step size:

$$\|x^{k+1} - x^*\| \leq \sqrt{1 - \frac{1}{\kappa^2}} \|x^k - x^*\|$$

The number of iterations needed for ϵ-accuracy is $O(\kappa^2 \log(1/\epsilon))$.

Theorem 13 (Convergence for General Convex Functions)

Assume:

1. f *has* L*-Lipschitz gradient on* Ω
2. f *is convex (but not necessarily strongly convex)*
3. *Step size satisfies* $0 < \alpha \leq \frac{1}{L}$

Then the gradient projection method satisfies:

1. $f(x^{k+1}) \leq f(x^k) - \frac{1}{2\alpha}\|x^{k+1} - x^k\|^2$
2. $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$
3. *Every limit point is a stationary point*
4. *If solution set is nonempty,* $\{f(x^k)\}$ *converges to the optimal value*

Proof.

Using the Lipschitz gradient and projection properties, one can show:

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2$$

From the projection optimality condition:

$$(x^{k+1} - x^k + \alpha \nabla f(x^k))^\top (x - x^{k+1}) \geq 0, \quad \forall x \in \Omega$$

Taking $x = x^k$ and combining inequalities yields the sufficient decrease property. The convergence results follow by standard arguments. □

Contents

Convex set constrained problem

$$\min_x \quad f(x), \quad \text{s.t. } x \in \Omega.$$

- f is convex and differentiable
- C is closed bounded and convex, e.g., $C = \{x \mid \|x\|_p \leq R\}$

Construct local approximation at x^k:

$$\begin{cases} \min & f(x^k) + \nabla f(x^k)^T (x - x^k) \\ \text{s.t.} & x \in \Omega. \end{cases}$$

Interpretation:

- Linear approximation of f but don't go outside

Frank-Wolfe Algorithm/Conditional Gradient Method:

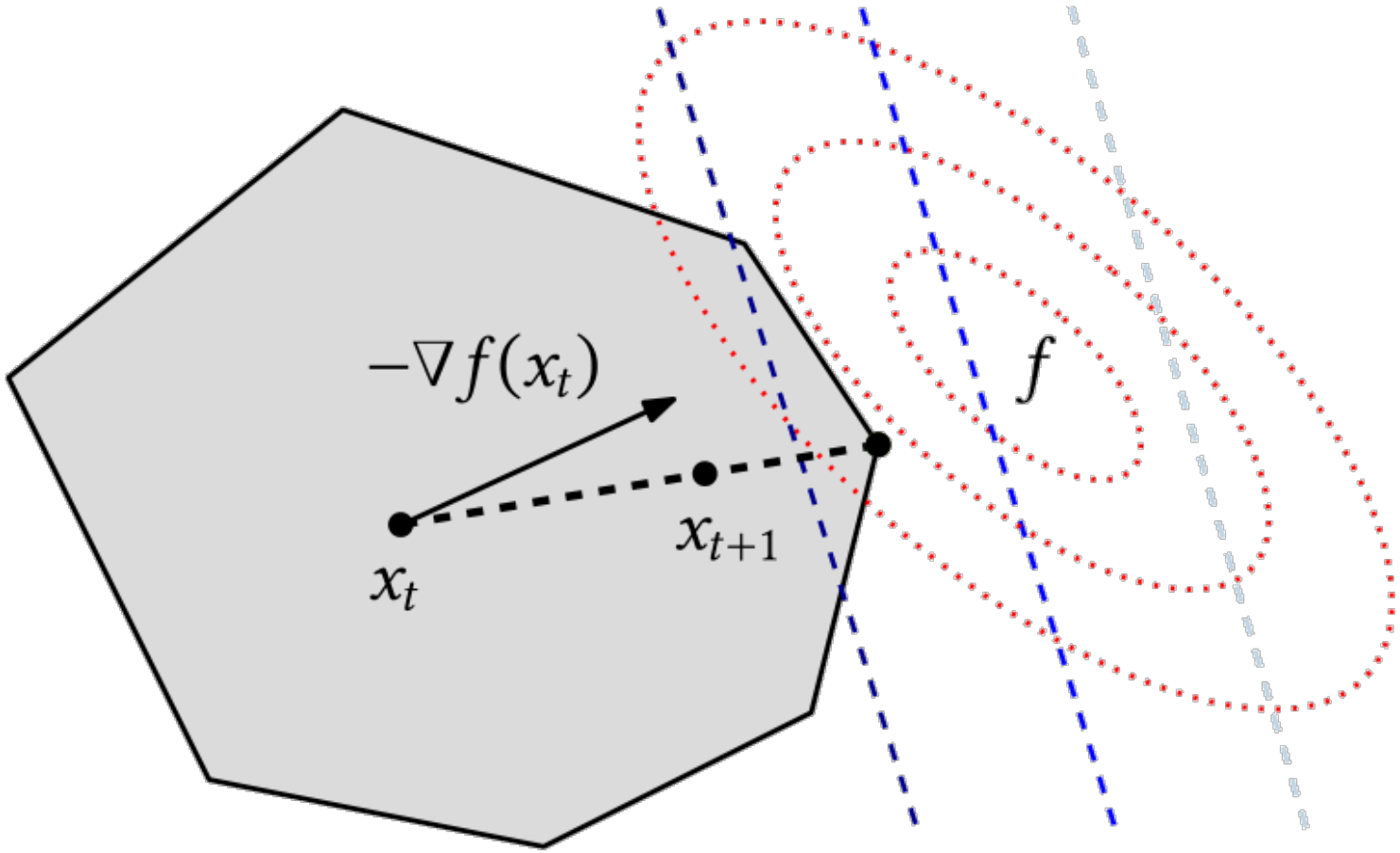
$$s^k \in \arg\min_{s \in \Omega} \nabla f(x^k)^T s$$

$$x^{k+1} \leftarrow (1 - \gamma_k) x^k + \gamma_k s^k$$

Stepsizes: $\gamma_k = 2/(k+2), k = 1, 2, ...$ Or, line search like backtracking Armijo.

Note for $\gamma_k \in (0, 1)$, we have $x^{k+1} \in \Omega$ by convexity. Can rewrite update as

$$x^{k+1} \leftarrow x^k + \gamma_k (s^k - x^k)$$



$$f(x_t) + \langle \nabla f(x_t), x - x_t \rangle$$

Convergence without line search

Let L be the L-Lipschitz constant

$$
\begin{aligned}
f(x^{k+1}) &= f((1-\gamma_k)x^k + \gamma_k s^k) \\
&\le f(x^k) + \gamma_k \langle s^k - x^k, \nabla f(x^k)\rangle + \frac{L}{2}\gamma_k^2 \|s^k - x^k\|_2^2 \\
&\le f(x^k) + \gamma_k \langle x^* - x^k, \nabla f(x^k)\rangle + \frac{L}{2}\gamma_k^2 \|s^k - x^k\|_2^2 \\
&= (1-\gamma_k) f(x^k) + \gamma_k \underbrace{(f(x^k) + \langle x^* - x^k, \nabla f(x^k)\rangle)}_{\text{convexity}} + \frac{L}{2}\gamma_k^2 \|s^k - x^k\|_2^2 \\
&\le (1-\gamma_k) f(x^k) + \gamma_k f(x^*) + \frac{1}{2}\gamma_k^2 L D^2
\end{aligned}
$$

Hence,

$$
f(x^{k+1}) - f(x^*) \le (1-\gamma_k)(f(x^k) - f(x^*)) + \frac{LD^2}{2}\gamma_k^2
$$

By choosing $\gamma_k = \frac{2}{k+2}$, it follows from induction that

$$
f(x^k) - f(x^*) \le (1 - \frac{2}{k+2})\frac{2LD^2}{k+1} + \frac{LD^2}{2}(\frac{2}{k+2})^2 \le \frac{2LD^2}{k+2}.
$$

Nonconvex cases: combine line search and apply Zoutendjik's results to get

$$\frac{(\nabla f(x^k)^T d^k)^2}{\|d^k\|^2} \to 0$$

$\|d^k\| \leq D$, it follows that

$$\min_{x \in \Omega} \langle \nabla f(x^k), x - x^k \rangle \to 0$$

So that for every limit point:

$$\min_{x \in \Omega} \langle \nabla f(x^*), x - x^* \rangle = 0$$

Or,

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

This means,

$$\nabla f(x^*) \in \mathcal{N}_\Omega(x^*)$$