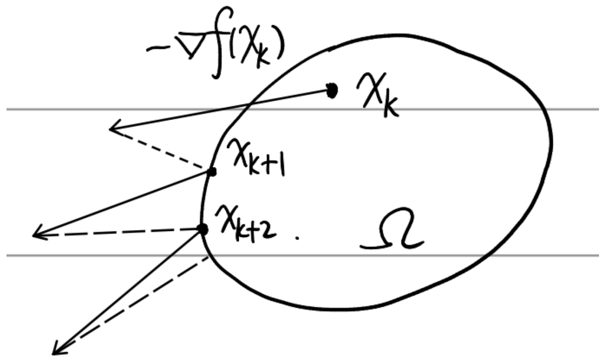
Lecture 17. Projected Gradient Descent

17.1 Projection operator and projected gradient descent

To solve the inequality constrained problems, we introduce the *projected gradient descent*.

Recall the iteration step in the gradient descent method, $x_{k+1} = x_k - \eta \nabla f(x_k)$. Now we need to minimize $f(x)$ over a feasible set Ω. If $x_k - \eta \nabla f(x_k)$ is feasible, then we can run the gradient descent iteration. If $x_k - \eta \nabla f(x_k)$ is infeasible, a simple idea is to project it onto Ω. This method is called the *projected gradient descent*.



Definition (*Projection*)

The projection of a point onto a set is the point in the set with minimum distance to the given point. Namely, the *projection operator* is defined by

$$\mathcal{P}_\Omega(\boldsymbol{y}) = \arg\min_{\boldsymbol{x} \in \Omega} \|\boldsymbol{x} - \boldsymbol{y}\| .$$

The the projected gradient descent step can be given by

$$\boldsymbol{x}_{k+1} = \mathcal{P}_\Omega\big(\boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)\big) .$$

Let

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{1}{\eta}\Big(\boldsymbol{x} - \mathcal{P}_\Omega\big(\boldsymbol{x} - \eta \nabla f(\boldsymbol{x})\big)\Big) ,$$

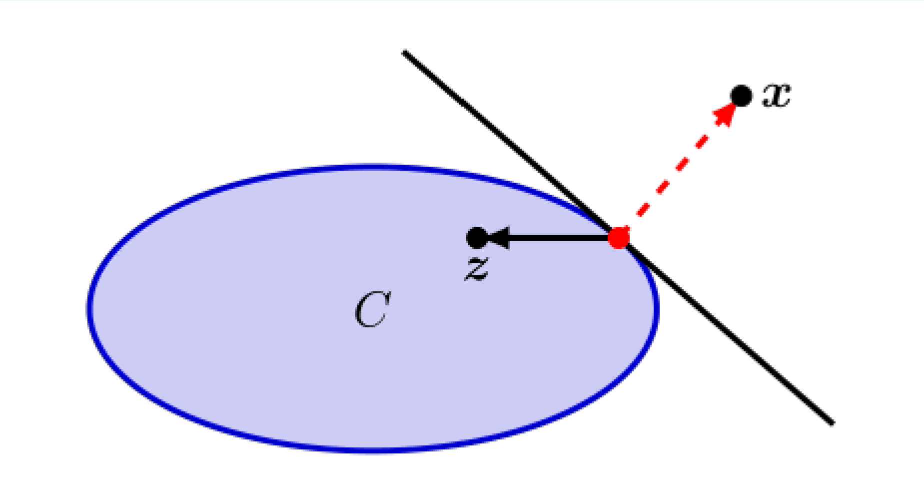
the iteration step can be expressed as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta \boldsymbol{g}(\boldsymbol{x}_k) \, .$$

Recall that, in Lecture 4, we show the following lemma.

Lemma

Let C be a nonempty, closed and convex set. Given $\boldsymbol{x}$ and $\boldsymbol{y} = \mathcal{P}_C(\boldsymbol{x})$, for any $\boldsymbol{z} \in C$, it holds that $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{y} \rangle \leq 0$.



Conversely, if there exists $\boldsymbol{y} \in C$ such that $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{y} \rangle \leq 0$, we have $\boldsymbol{y} = \mathcal{P}_C(\boldsymbol{x})$. Otherwise, let $\boldsymbol{w} = \mathcal{P}_C(\boldsymbol{x})$. Then we have

$$\langle \boldsymbol{x} - \boldsymbol{w}, \boldsymbol{y} - \boldsymbol{w} \rangle \leq 0 \, .$$

However, we also have $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{w} - \boldsymbol{y} \rangle \leq 0$, which implies that

$$\langle \boldsymbol{x} - \boldsymbol{w}, \boldsymbol{w} - \boldsymbol{y} \rangle = \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{w} - \boldsymbol{y} \rangle + \langle \boldsymbol{y} - \boldsymbol{w}, \boldsymbol{w} - \boldsymbol{y} \rangle < 0$$

if $\boldsymbol{y} \neq \boldsymbol{w}$. Contradiction.
Thus, $\boldsymbol{y} = \mathcal{P}_C(\boldsymbol{x})$ if and only if $\langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{z} - \boldsymbol{y} \rangle$ for any $\boldsymbol{z} \in C$.

Applying this lemma, we can show that $\boldsymbol{g}(\boldsymbol{x})$ plays a similar role as $\nabla f(\boldsymbol{x})$ in the gradient descent.

Lemma

For any $\boldsymbol{x} \in \Omega$,

$$\langle \nabla f(\boldsymbol{x}),\, \boldsymbol{g}(\boldsymbol{x}) \rangle \geq 0\,.$$

The inequality holds if and only if $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$.

Proof

Since $\boldsymbol{x} \in \Omega$, we have

$$\langle \boldsymbol{x} - \mathcal{P}_\Omega(\boldsymbol{x} - \eta \nabla f(\boldsymbol{x})), \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}) - \mathcal{P}_\Omega(\boldsymbol{x} - \eta \nabla f(\boldsymbol{x})) \rangle \leq 0\,,$$

which gives that

$$\langle \eta \boldsymbol{g}(\boldsymbol{x}), \eta \boldsymbol{g}(\boldsymbol{x}) - \eta \nabla f(\boldsymbol{x}) \rangle = \eta^2 \langle \boldsymbol{g}(\boldsymbol{x}),\, \boldsymbol{g}(\boldsymbol{x}) - \nabla f(\boldsymbol{x}) \rangle \leq 0\,.$$

Thus,

$$\langle \nabla f(\boldsymbol{x}),\, \boldsymbol{g}(\boldsymbol{x}) \rangle \geq \langle \boldsymbol{g}(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x}) \rangle\,.$$

So we know that $-\boldsymbol{g}(\boldsymbol{x})$ is a desceding direction. Now we show that if $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$ then $\boldsymbol{x}$ is a minimum point.

Lemma

$\boldsymbol{x}^*$ is a minimum point of f over Ω, iff $\boldsymbol{g}(\boldsymbol{x}) = \mathbf{0}$, namely, $\boldsymbol{x}^* = \mathcal{P}_\Omega(\boldsymbol{x}^* - \eta \nabla f(\boldsymbol{x}^*))$.

Proof

Applying the above lemma, we have $\boldsymbol{x}^* = \mathcal{P}_\Omega(\boldsymbol{x}^* - \nabla f(\boldsymbol{x}^*))$ if and only if

$$\langle \boldsymbol{x}^* - \eta \nabla f(\boldsymbol{x}^*) - \boldsymbol{x}^*, \boldsymbol{z} - \boldsymbol{x}^* \rangle \leq 0$$

for all $\boldsymbol{z} \in \Omega$, which is further equivalent to

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

We conclude this lemma by the first-order optimality conditions of convex functions.

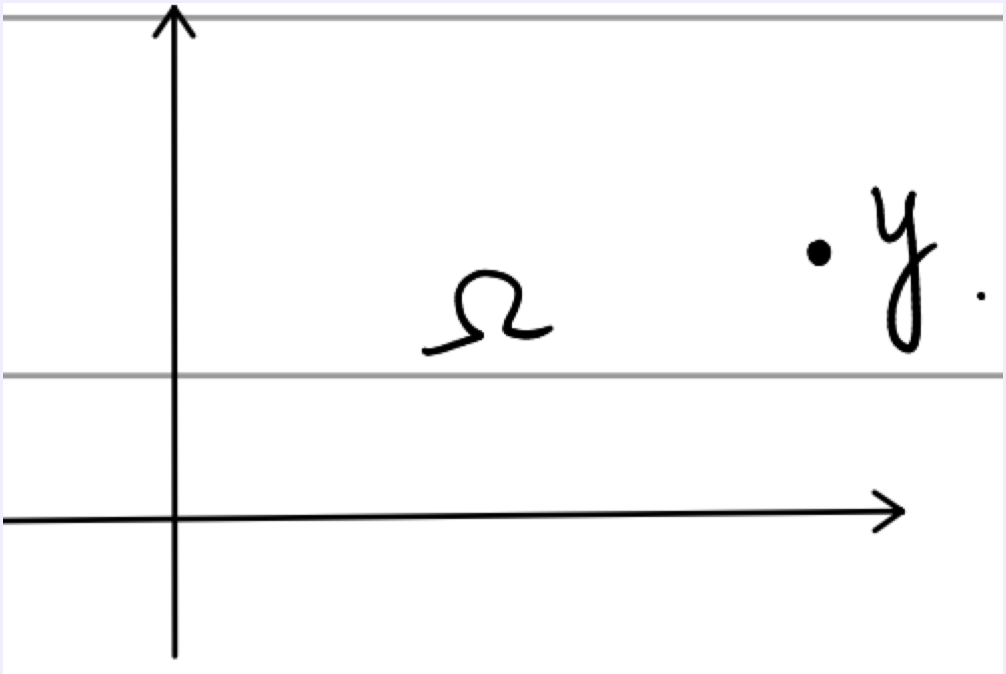
Hence, in the projected gradient descent, we can stop when $\boldsymbol{g}(\boldsymbol{x}_k)$ is small, or equivalently when $\boldsymbol{x}_{k+1} - \boldsymbol{x}_k$ is small.

17.2 Examples of projection operator

Projected gradient descent is useful when the projection operator can be computed efficiently. Here we give some examples.

Example 1 (*Box constraints*)

$$\Omega = \{x \mid a_i \leq x_i \leq b_i, \quad i = 1, \cdots, n\}$$

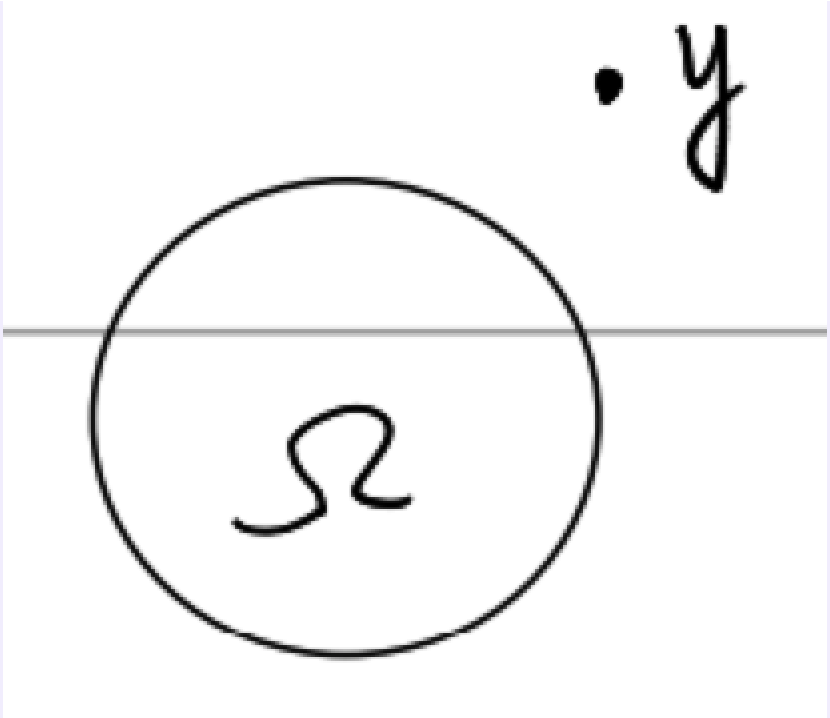


It is easy to see that

$$[\mathcal{P}_\Omega(y)]_i = \min\{b_i, \max\{a_i, y_i\}\} = \begin{cases} a_i & y_i < a_i \\ y_i & a_i \leq y_i \leq b_i \\ b_i & y_i > b_i \end{cases}$$

Example 2 (*L^2 constraints, ridge regression*)

$$\Omega = \{x \mid \|x\|_2 \leq t\}$$



The projection operator $\mathcal{P}_\Omega(y)$ is to compute

$$\begin{aligned} \min \quad & \|x-y\|^2 \\ \text{subject to} \quad & \|x\|_2^2 \le t^2 \end{aligned}$$

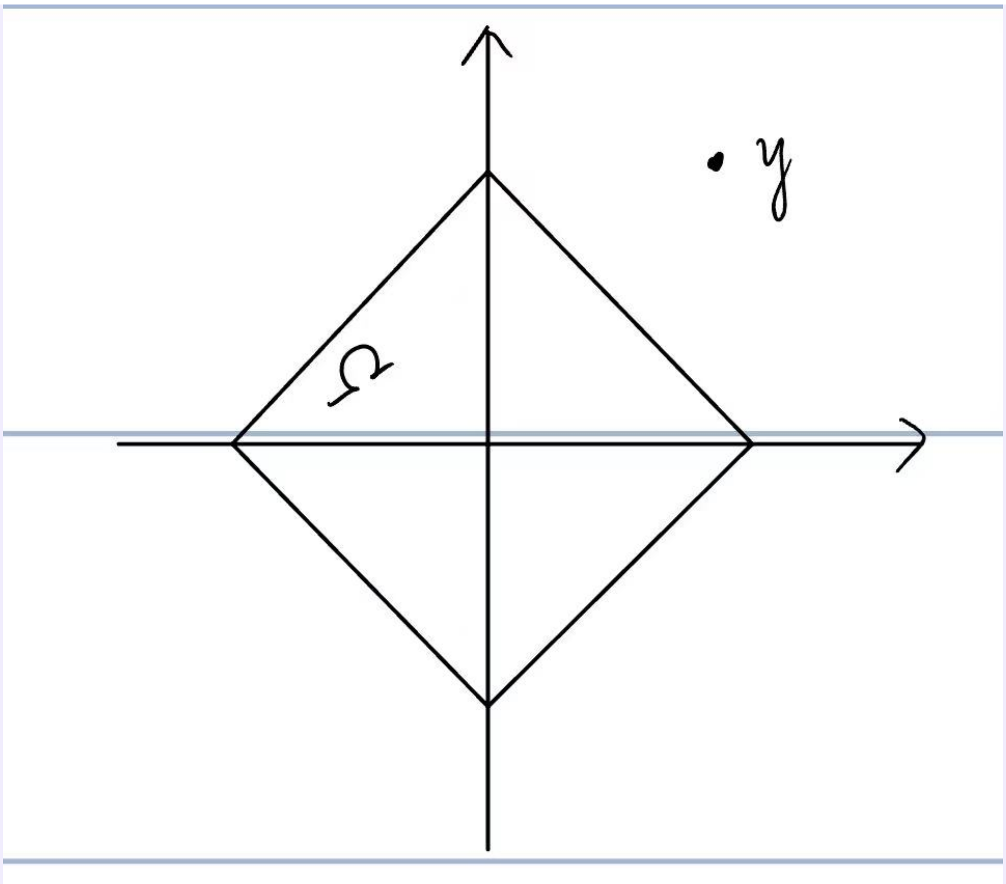
By KKT condition, there exists $\mu \ge 0$ such that

$$2(x-y) + 2\mu x = 0 \quad \text{and} \quad \mu(\|x\|^2 - t) = 0$$

Then we have $y = (1+\mu)x$.
Hence, $\mathcal{P}_\Omega(y) = \min\left\{1, \frac{t}{\|y\|_2}\right\} y$.

Example 3 (*L^1 constraints, LASSO*)

$$\Omega = \{x : \|x\|_1 \le t\}$$



Unfortunately, there is no closed form for the projection operator $\mathcal{P}_\Omega(y)$. But we can compute it efficiently.
By symmetry, we only need to consider the case where $y_i \geq 0$ for all i. Now $\mathcal{P}_\Omega(y)$ is equivalent to the following optimization problem:

$$\begin{aligned} \min \quad & \|x-y\|^2 \\ \text{subject to} \quad & \sum_i x_i \leq t \\ & x_i \geq 0, \forall\, i\,. \end{aligned}$$

By KKT condition, assume there exist KKT multipliers $\mu_0, \cdots, \mu_n$ such that

$$\begin{cases} 2(x_i - y_i) + \mu_0 - \mu_i = 0, \forall i \\ \mu_0(\sum x_i - t) = 0 \\ \mu_i x_i = 0 \\ \mu_i \geq 0 \\ \sum x_i \leq t, x_i \geq 0 \end{cases}$$

- Case 1. $\|y\|_1 \leq t$, then $\mu_0 = \mu_i = 0$. Hence $x = y$.
- Case 2. $\|y\|_1 > t$, then $\sum 2(x_i - y_i) + \mu_0 - \mu_1 = 2(\sum x_i - \sum y_i) + n\mu_0 - \sum \mu_i = 0$, hence $\mu_0 > 0$. Since $\mu_0(\sum x_i - t) = 0$, we have $\sum x_i = t$.
  - If $\mu_i = 0$, by $2(x_i - y_i) + \mu_0 - \mu_i = 0$, we have $x_i = y_i - \frac{1}{2}\mu_0$.

- If $\mu_i > 0$, by $\mu_i x_i = 0$, we have $x_i = 0$.

Now we have

$$x_i = \begin{cases} y_i - \frac{1}{2}\mu_0 & \text{if } y_i \geq \frac{1}{2}\mu_0 \\ 0 & \text{otherwise} \end{cases}$$

and $\sum x_i = t$.
We may use the binary search to find μ_0, where the lower bound is 0 and the upper bound is $\max y_i$.

17.3 Comparison with proximal gradient descent

To analyze the convergence of the projected gradient descent, we show that it is a special case of the proximal gradient descent.

Let I_Ω be the *indicator function* of Ω, defined by

$$I_\Omega(\boldsymbol{x}) = \begin{cases} 0 & \boldsymbol{x} \in \Omega \\ \infty & \boldsymbol{x} \notin \Omega \end{cases}.$$

Clearly I_Ω is a convex function if and only if Ω is a convex set.

Then we can show that the proximal operator for I_Ω is simply the projection onto Ω:

$$\begin{aligned} \operatorname{prox}_{I_\Omega}(\boldsymbol{y}) &= \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2 + I_\Omega(\boldsymbol{x}) \\ &= \arg\min_{\boldsymbol{x}\in\Omega} \|\boldsymbol{x} - \boldsymbol{y}\|^2 \\ &= \mathcal{P}_\Omega(\boldsymbol{y}) \,. \end{aligned}$$

Since

$$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \quad \Longleftrightarrow \quad \min_{\boldsymbol{x}} f(\boldsymbol{x}) + I_\Omega \boldsymbol{x} \,,$$

and for any $\eta > 0$,

$$\boldsymbol{x}_{k+1} = \mathcal{P}_\Omega(\boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)) = \operatorname{prox}_{I_\Omega}(\boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)) = \operatorname{prox}_{\eta I_\Omega}(\boldsymbol{x}_k - \eta \nabla f(\boldsymbol{x}_k)) \,,$$

we find that the projected gradient descent for $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$ is the same as proximal gradient descent for $\min_{\boldsymbol{x}} f(\boldsymbol{x}) + I_\Omega(\boldsymbol{x})$.
By extending the results on Lecture 13 of to

$\varphi(\boldsymbol{x}) = f(\boldsymbol{x}) + I_\Omega(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the convergence analysis for proximal gradient descent applies also to projected gradient descent.

Theorem

Let Ω be a nonempty convex set, and f be an L-smooth convex function over Ω. Suppose $\boldsymbol{x}^*$ is a minimum of f over Ω. Then the sequence $\{\boldsymbol{x}_k\}$ produced by projected gradient descent with constant step size $\eta \in (0, 1/L]$ satisfies $f(\boldsymbol{x}_{k+1}) \leq f(\boldsymbol{x}_k)$ and

$$f(\boldsymbol{x}_k) - f(\boldsymbol{x}^*) \leq \frac{\|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2}{2\eta k}\,.$$

Furthermore, if f is also μ-strongly convex, then

$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|^2 \leq (1 - \mu\eta)^k \|\boldsymbol{x}^* - \boldsymbol{x}_0\|^2\,.$$