

Optimization for Machine Learning

机器学习中的优化方法

陈程

华东师范大学 软件工程学院

chchen@sei.ecnu.edu.cn

Review of Gradient Descent

Consider an unconstrained convex optimization problem

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}).$$

The **gradient descent** method starts with an initial point \mathbf{x}_0 , and iteratively computes

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t).$$

Review of Gradient Descent

	stepsize	convergence rate	iteration complexity
strongly convex & smooth	$\eta_t = \frac{1}{L}$ or $\eta_t = \frac{2}{\mu+L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O\left(\kappa \log \frac{1}{\varepsilon}\right)$
locally strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O\left(\kappa \log \frac{1}{\varepsilon}\right)$
PL condition & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O\left(\kappa \log \frac{1}{\varepsilon}\right)$
convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\varepsilon}\right)$

Table: Convergence Property of GD

Outline

1 Projected gradient descent

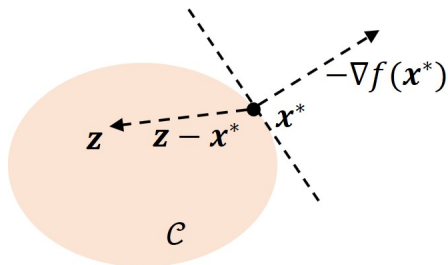
2 Frank-Wolfe algorithm

Constrained convex optimization

Suppose f is a **convex** function and $\mathcal{C} \in \mathbb{R}^d$ is a **closed** and **convex** set. The constrained convex optimization problem is:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{C} \end{aligned}$$

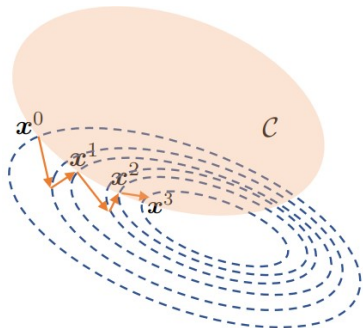
Optimality condition



Suppose f is convex and differentiable, \mathcal{C} is close and convex. Then

$$\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x}) \iff \langle -\nabla f(\mathbf{x}^*), \mathbf{z} - \mathbf{x}^* \rangle \leq 0, \forall \mathbf{z} \in \mathcal{C}$$

Projected gradient descent (投影梯度下降法)



Idea: project onto \mathcal{C} after every gradient descent step:

$$\mathbf{x}_{t+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)).$$

where $\mathcal{P}_{\mathcal{C}}(\mathbf{x}) \triangleq \arg \min_{\mathbf{z} \in \mathcal{C}} \|\mathbf{z} - \mathbf{x}\|_2^2$ is Euclidean projection onto \mathcal{C} .

Think

Suppose f is a convex function and \mathcal{C} is a closed convex set. Let

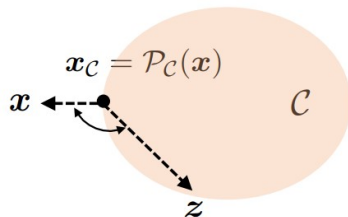
$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}) \quad \text{and} \quad \mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$$

Is it true that

$$\mathbf{x}^* = \mathcal{P}_{\mathcal{C}}(\hat{\mathbf{x}})?$$

The answer is NO.

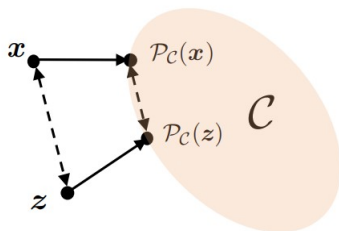
Properties of projection



Let $C \in \mathbb{R}^d$ be closed and convex, $\mathbf{z} \in C$, $\mathbf{x} \in \mathbb{R}^d$. Then

- $\langle \mathbf{x} - \mathcal{P}_C(\mathbf{x}), \mathbf{z} - \mathcal{P}_C(\mathbf{x}) \rangle \leq 0$.
- $\|\mathbf{x} - \mathcal{P}_C(\mathbf{x})\|_2^2 + \|\mathbf{z} - \mathcal{P}_C(\mathbf{x})\|_2^2 \leq \|\mathbf{x} - \mathbf{z}\|_2^2$

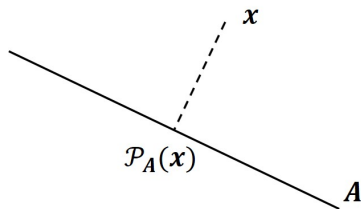
Properties of projection: nonexpansiveness



Let $\mathcal{C} \in \mathbb{R}^d$ be closed and convex. For any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, we have

$$\|\mathcal{P}_{\mathcal{C}}(\mathbf{x}) - \mathcal{P}_{\mathcal{C}}(\mathbf{z})\|_2 \leq \|\mathbf{x} - \mathbf{z}\|_2.$$

Examples

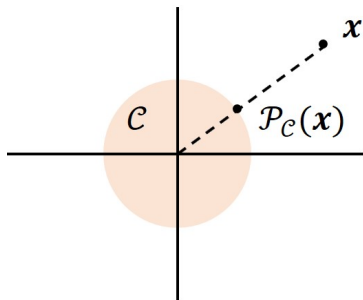


Projecting onto an affine subspace:

$$\mathbf{y} = \arg \min_z \|\mathbf{A}z - \mathbf{x}\|_2 = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{x}$$

$$\mathcal{P}_A(\mathbf{x}) = \mathbf{A}\mathbf{y} = \mathbf{A}(\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{x}$$

Examples



Projecting onto a unit Euclidean ball (ℓ_2 ball):

$$\mathcal{P}_C(\mathbf{x}) = \arg \min_{\|\mathbf{z}\|_2 \leq 1} \|\mathbf{x} - \mathbf{z}\|_2 = \frac{\mathbf{x}}{\max\{1, \|\mathbf{x}\|_2\}}$$

Smooth and strongly convex constrained problems

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{C} \end{aligned}$$

- f : L -smooth and μ -strongly convex
- $\mathcal{C} \in \mathbb{R}^d$: closed and convex

Smooth and strongly convex constrained problems

Let f be L -smooth and μ -strongly convex. If $\eta_t = \eta = \frac{2}{\mu+L}$, then PGD obeys

$$\|\mathbf{x}_t - \mathbf{x}^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right)^t \|\mathbf{x}_0 - \mathbf{x}^*\|_2.$$

the same convergence rate as for the unconstrained case!

Contraction Mapping (压缩映射)

Contraction mapping in Euclidean space: If a function $f : \mathcal{X} \rightarrow \mathcal{X}$ satisfies

$$\|f(x) - f(y)\|_2 \leq \gamma \|x - y\|_2, \quad \forall x, y \in \mathcal{X}$$

for some $\gamma \in (0, 1)$, then we call f is a contraction mapping.

The contraction mapping f has a unique fixed point \hat{x} , i.e., $f(\hat{x}) = \hat{x}$.

Smooth and convex constrained problems

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathcal{C} \end{aligned}$$

- f : **convex** and L -smooth
- $\mathcal{C} \in \mathbb{R}^d$: closed and convex

Smooth and convex constrained problems

Let f be convex and L -smooth. If $\eta_t = \eta = \frac{1}{L}$, then PGD obeys

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{t}$$

the same convergence rate as for the unconstrained case

Convergence analysis

Recall the main steps when handling the unconstrained case:

- **Step 1:** show improvement

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|_2^2 \quad \text{not true for constrained case}$$

- **Step 2:** by convexity,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}^*) - \langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \rangle - \frac{1}{2L} \|\nabla f(\mathbf{x}_t)\|_2^2 \\ &= f(\mathbf{x}^*) + \frac{L}{2} \left\{ \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \left\| \mathbf{x}_t - \mathbf{x}^* - \frac{1}{L} \nabla f(\mathbf{x}_t) \right\|_2^2 \right\} \end{aligned}$$

- **Step 3:** telescoping

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} (f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)) \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Convergence analysis

For the constrained case, we aim to replace $\nabla f(\mathbf{x})$ in the unconstrained case by

$$g_C(\mathbf{x}) = L(\mathbf{x} - \mathcal{P}_C(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})))$$

We have $g_C(\mathbf{x}_t) = L(\mathbf{x}_t - \mathbf{x}_{t+1})$.

- **Step 1:** descent guarantee

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L} \|g_C(\mathbf{x}_t)\|_2^2$$

- **Step 2:**

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}^*) - \langle g_C(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t \rangle - \frac{1}{2L} \|g_C(\mathbf{x}_t)\|_2^2$$

- **Step 3:** telescoping

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{1}{T} \sum_{t=0}^{T-1} (f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*)) \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^*\|^2$$

Outline

1 Projected gradient descent

2 Frank-Wolfe algorithm

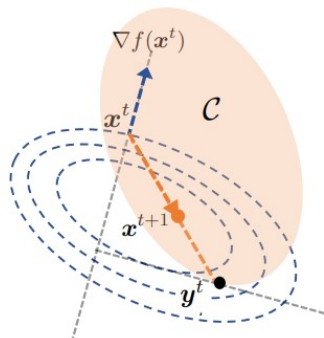
Frank-Wolfe Algorithm

Consider following problem:

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{Ax} \leq \mathbf{b} \end{aligned}$$

Computing projection is very expensive!

Frank-Wolfe Algorithm



Algorithm 1 Frank-Wolfe (a.k.a. conditional gradient) Algorithm

for $t = 1, 2, \dots$ **do**

$\mathbf{y}_t = \arg \min_{\mathbf{x} \in C} \langle \nabla f(\mathbf{x}_t), \mathbf{x} \rangle$ //direction finding

$\mathbf{x}_{t+1} = (1 - \eta_t)\mathbf{x}_t + \eta_t\mathbf{y}_t$ //line search and update

Frank-Wolfe Algorithm

Algorithm 2 Frank-Wolfe (a.k.a. conditional gradient) Algorithm

for $t = 1, 2, \dots$ **do**

$\mathbf{y}_t = \arg \min_{\mathbf{x} \in \mathcal{C}} \langle \nabla f(\mathbf{x}_t), \mathbf{x} \rangle$ //direction finding

$\mathbf{x}_{t+1} = (1 - \eta_t)\mathbf{x}_t + \eta_t\mathbf{y}_t$ //line search and update

- main step: linearization of the objective function
- appealing when linear optimization is much cheaper than projection
- stepsize: $\eta_t = \frac{2}{t+2}$

Frank-Wolfe Algorithm

Let f be convex and L -smooth. If $\eta_t = \frac{2}{t+2}$, one has

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \frac{2LD^2}{t+2}$$

where $D = \sup_{\mathbf{x}, \mathbf{y} \in \mathcal{C}} \|\mathbf{x} - \mathbf{y}\|_2$

For **compact** constraint sets, Frank-Wolfe attains ε -accuracy with $O(\frac{1}{\varepsilon})$ iterations.

Summary

Table: Convergence Property of Projected GD

	stepsize	convergence rate	iteration complexity
strongly convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\left(1 - \frac{1}{\kappa}\right)^t\right)$	$O(\kappa \log \frac{1}{\epsilon})$
convex & smooth	$\eta_t = \frac{1}{L}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\epsilon}\right)$

Table: Convergence Property of FW

	stepsize	convergence rate	iteration complexity
convex & smooth	$\eta_t = \frac{1}{t}$	$O\left(\frac{1}{t}\right)$	$O\left(\frac{1}{\epsilon}\right)$