Chapter 8: Optimization for Data Science
Projected Gradient Descent

TANN Chantara

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

October 9, 2022

Table of Contents

Constrained Optimization Problems

Constrained minimization problem

$$\min_{x \in \mathcal{X}} f(x) = f(x^*)$$

- $\mathcal{X} \subseteq \mathbb{R}^n$ closed and convex.
- $f : \mathbb{R}^n \to \mathbb{R}$ convex and differentiable

Goal: Find a approximate solution $\in \mathbb{R}^n$ such that

$$f(\tilde{x}) - f(x^*) < \varepsilon$$

- Compute iteratively via projected gradient descent.

Projected Gradient Descent Algorithm

We choose an arbitrary $x_0 \in \mathcal{X}$ and for $t > 0$ define

Projected gradient descent:

$$
\begin{aligned}
y_{t+1} &= x_t - \gamma \nabla f(x_t) \\
x_{t+1} &= \Pi_{\mathcal{X}}(y_{t+1}) = \arg\min_{x \in \mathcal{X}} ||x - y_{t+1}||^2
\end{aligned}
$$

- Projection onto $\mathcal{X}$ ensures $x_t \in \mathcal{X}$ for all $t > 0$.
- Projection is well-defined as $||x - y||^2$ is strongly convex in x.
- Computing $\Pi_{\mathcal{X}}(y_{t+1})$ means to solve an auxiliary convex constrained minimization problem in each step.

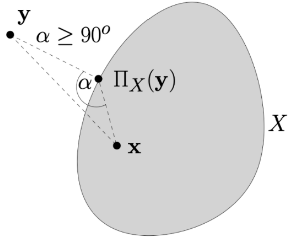
Auxiliary Results on the Projection

Facts: For every $x \in \mathcal{X}, y \in \mathbb{R}^n$

(i) $(x - \Pi_{\mathcal{X}}(y))^T(y - \Pi_{\mathcal{X}}(y)) \leq 0$

(ii) $||x - \Pi_{\mathcal{X}}(y)||^2 + ||y - \Pi_{\mathcal{X}}(y)||^2 \leq ||x - y||^2$

Proof: To show (i), recall that optimization conditions for convex problems state that $\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in \mathcal{X}$. We now consider $f(x) = ||x - y||^2$ and let $x^* = \min_{x \in \mathcal{X}} f(x) = \Pi_{\mathcal{X}}(y)$. Then (i) follows directly. Assertion (ii) follows from (i) via the equation $2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$, which holds for any $v, w \in \mathbb{R}^n$.



Bounded Gradients

Theorem (Projected gradient descent for Lipschitz functions:)
Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable and $\mathcal{X} \subset \mathbb{R}^n$ closed, convex with global minimum $x^* \in \mathcal{X}$. Suppose that $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B, \forall x \in \mathbb{R}^n$. Choosing the step size $\gamma = R/B\sqrt{T}$, the projected gradient descent for $x_0 \in \mathcal{X}$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{RB}{\sqrt{T}}$$

Smooth Convex Functions

Lemma (Sufficient decrease): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, L-smooth and $\mathcal{X} \subset \mathbb{R}^n$ be closed, convex. For $\gamma = 1/L$, projected gradient descent with any $x_0 \in \mathcal{X}$ yields

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||\nabla f(x_t)||^2 + \frac{L}{2}||y_{t+1} - x_{t+1}||^2, t \geq 0$$

Theorem (Projected gradient descent for smooth functions): Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex, differentiable, L-smooth and $\mathcal{X} \in \mathbb{R}^n$ be closed, convex with global minimum x^*. For $\gamma = 1/L$ and any $x - 0 \in \mathcal{X}$ projected gradient descent yields

$$f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2, T > 0$$

Smooth and Strongly Convex Functions

Theorem (Smooth and Strongly Convex Function): Let $f : \mathbb{R}^n \to \mathbb{R}$ be μ-strongly convex, differentiable, L-smooth and $\mathcal{X} \subset \mathbb{R}^n$ be closed, convex with global minimum x^*. For a step size $\gamma = 1/L$ and any $x_0 \in \mathcal{X}$, projected gradient descent yields

(i) $||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2, t \geq 0.$

(ii) $$f(x_T) - f(x^*) \leq ||\nabla f(x^*)||(1 - \tfrac{\mu}{L})^{T/2}||x_0 - x^*||^2 + \frac{L}{2}(1 - \tfrac{\mu}{L})^T||x_0 - x^*||^2, T > 0$$

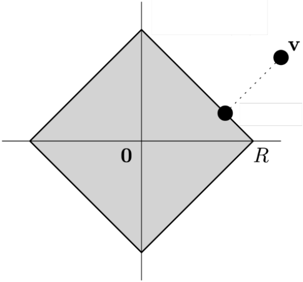
- Recall that $\nabla f(x^*)$ does not necessarily vanish in the constrained case.
- Given again an iteration complexity of $\mathcal{O}(\log(1/\varepsilon))$

Projecting onto l_1 balls

Goal: Compute $\Pi_{\mathcal{X}}(v)$

$$\mathcal{X} = \underbrace{\left\{ x \in \mathbb{R}^n : ||x||_1 = \sum_{i=1}^{d} |x_i| \leq R \right\}}_{=\mathbb{B}_1(R)}$$

- $\mathcal{X}$ is a polytope with 2^n many facets.
- Problem can be simplified in several steps



Fact 1: We can assume without loss of generality that (i) $R = 1$, (ii) $v_i \geq 0$ for all i and (iii) $\sum_{i=1}^{n} v_i > 1$.

Proof of Fact 1

(i) If we project v/R onto $\mathbb{B}_1(1)$, we obtain $\Pi_{\mathcal{X}}(v)/R$, so we can restrict to $R = 1$.

(ii) Observe that simultaneously flipping the signs of a fixed subset of coordinates in both v and $x \in \mathcal{X}$ yields vectors v' and $x' \in \mathcal{X}$ such that $||x - v|| = ||x' - v'||$; thus x minimizes the distance to v if and only if x' minimizes the distance to v'. Hence, it suffices to compute $\Pi_{\mathcal{X}}(v)$ for vectors with nonnegative entries.

(iii) If $\sum_{i=1}^{n} v_i \leq 1$, then $\Pi_{\mathcal{X}}(v) = v$ and there is nothing to compute, so the interesting case is $\sum_{i=1}^{n} v_i > 1$

Projecting onto l_1 balls Cont'd

Fact 2: Under the assumptions of Fact 1, $x = \Pi_{\mathcal{X}}(v)$ satisfies (i) $x_i \geq 0$ for all i and (ii) $\sum_{i=1}^{n} x_i = 1$.

Proof:

(i) Consider $x = \Pi_{\mathcal{X}}(v)$ and suppose $x_i < 0$ for some i. We show this leads to a contradiction and hence $x_i \geq 0$. Suppose $x_i < 0$ for some i, then $(-x_i - v_i) \leq (x_i - v_i)^2$, since $v_i \geq 0$. Therefore, flipping the sign of the $i-$th component of x would yield another vector in $\mathcal{X}$ at least as close to v as x. Since $x = \Pi_{\mathcal{X}}(v)$ and the 2-norm is strictly convex, this is impossible.

(ii) Suppose for the sake of contradiction that $\sum_{i=1}^{n} x_i < 1$, then considering $x' = x + \lambda(v - x) \in \mathcal{X}$ for some small $\lambda > 0$, but then $||x' - v|| = (1 - \lambda)||x - v|| < ||x - v||$, which contradicts the optimality of x. Hence, $x = \Pi_{\mathcal{X}}(v) = \arg\min_{x \in \Delta_n} ||x - v||^2$

Projecting onto l_1 balls Cont'd

Fact 3: We assume without loss of generality that $v_1 \geq \cdots \geq v_n$.

Lemma 1: Let $x^* = \arg\min_{x \in \Delta_n} ||x - v||^2$. Under assumption of Fact 3, there exists a unique $p \in \{1, ..., n\}$ such that

$$\begin{cases} x_i^* > 0, & i \leq p \\ x_i^* = 0, & i > p \end{cases}$$

Proof: We consider the convex function $d_v(z) = ||z - v||^2$ and by the optimality criterion

$$\nabla d_v(x^*)(x - x^*) = 2(x^* - v)^T(x - x^*) \geq 0, x \in \Delta_n \qquad (1)$$

Projecting onto l_1 balls cont'd

Lemma 2: Let $x^* = \arg\min_{x \in \Delta_n} ||x - v||^2$. Under assumption of Fact 3 and with p as in Lemma 1

$$x_i^* = v_i - \theta_p, i \leq p, \quad \text{where} \quad \theta_p = \frac{1}{p}(\sum_{i=1}^{p} v_i - 1)$$

Proof: We argue by contradiction. If not all $x_i^* - v_i$ for $i \leq p$ have the same value $-\theta_p$, then we have $x_i^* - v_i < x_j^* - v_j$ for some $i, j \leq p$. We can decrease x_j^* by some small $\varepsilon > 0$ and simultaneously increase x_i^* by ε to obtain $x \in \Delta_n$ such that

$$(x^* - v)^T(x - x^*) = (0 - v_i)\varepsilon - (x_{i+1}^* - v_{i+1})\varepsilon = \varepsilon(\underbrace{v_{i+1} - v_i}_{\leq 0} - \underbrace{x_{i+1}^*}_{>0}) < 0$$

which contradicts (1). The expression for θ_p is obtained from (con'd next page)

Projecting onto l_1 balls cont'd

$$1 = \sum_{i=1}^{p} x_i^* = \sum_{i=1}^{p} (v_i - \theta_p) = \sum_{i=1}^{p} v_i - p\theta_p \qquad (2)$$

What we have found so far: we have n candidates for x^*, namely the vectors

$$x^*(p) = (v_1 - \theta_p, ..., v_p - \theta_p, 0, ..., 0), \;\; p \in \{1, ..., n\}$$

But which p should we select?

- By Lemma 1, $v_p - \theta_p > 0$
- We could just simply choose p that $||x^*(p) - v||^2$ is minimal
- But there is an even simpler criterion

Projecting onto l_1 balls cont'd

Lemma 3: Under assumption of Fact 3 with $x^*(p)$ as in (2) and

$$p^* = \max\{p \in \{1, ..., n\} : v_p - \frac{1}{p}\left(\sum_{i=1}^{p} v_i - 1\right) > 0$$

it holds that

$$\arg\min_{x \in \Delta_n} ||x - v||^2 = x^*(p^*).$$

This can be computed in $\mathcal{O}(n \log n)$