Modern Optimization

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

Outline

Remark 2.1

Let $\Omega \subset \mathbb{R}^d$ be a closed convex body, we are interested in the following optimization problems:

$$\min f_0(x)$$
$$\text{s.t. } x \in \Omega.$$

Definition 2.1

The **projected gradient descent** *is defined as the following sequence of rules*

$$y^{k+1} = x^k - \gamma_k \nabla f(x^k)$$
$$x^{k+1} = P_\Omega(y^{k+1}),$$

where P_Ω corresponds to the projection onto the convex set Ω (POCS: Projection Onto Convex Sets is not unrelated...).

Remark 2.2

Let us review the main ideas behind the projected gradient descent:

- The first step is a basic gradient step
- The second step corrects the gradient step if it reaches a point out of the feasible set
- γ_k the step size may or may not vary
- The projection step might not be cheap!:

$$P_\Omega(x) := \operatorname*{argmin}_{v \in \Omega} \|x - v\|.$$

Proposition 2.1 (Left as exercise)

Let Ω be a closed convex body. Then the projection P_Ω are well defined (and unique for all $x \in \mathbb{R}^d$).

Remark 2.3

Note that the first step assumes a point $x^0 \in \Omega$. This means simply doing a first projection on the input point (or even 0).

Exercise 2.1

Let $\Omega = B^2_{x^*}(R)$ be the ℓ^2 ball centered at a given point $x^* \in \mathbb{R}^d$ and of radius $R > 0$. What is $P_\Omega(x)$ for any $x \in \mathbb{R}^d$.

Lemma 1

Let $\Omega \subseteq \mathbb{R}^d$ *be a closed convex body. Let* $x \in \Omega$ *and* $y \in \mathbb{R}^d$*. It holds*

1. $\langle x - P_\Omega(y), y - P_\Omega(y) \rangle \leq 0.$
2. $\|x - P_\Omega(y)\|^2 + \|y - P_\Omega(y)\|^2 \leq \|x - y\|^2.$

Theorem 2.1

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ *be a convex differentiable function. Assume furthermore that* $\Omega \subseteq \mathrm{dom}(f)$ *is a closed convex subset,* x^* *is a minimizer of* f *over* Ω*,* $\|x^0 - x^*\| \leq R$ *for some* $R > 0$ *and* $x^0 \in \Omega$*. If the gradient of* f *is bounded:* $\|\nabla f(x)\| \leq B$ *for all* $x \in \Omega$*, then choosing a gradient step of*

$$\gamma := \frac{R}{B\sqrt{K}}$$

ensures that the iterates generated by the projected gradient descent starting at x^0 *satisfy*

$$\frac{1}{K} \sum_{k=0}^{K-1} \left(f(x^k) - f(x^*) \right) \leq \frac{RB}{\sqrt{K}}.$$

Lemma 2 (Descent direction of projected gradient)

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ *be a convex differentiable* L*-smooth function over a closed and convex set* $\Omega \subseteq \mathrm{dom}(f)$*. Given a constant stepsize*

$$\gamma = \frac{1}{L},$$

the sequence of iterates of the projected gradient, starting at $x^0 \in \Omega$ *satisfies*

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

Lemma 3

Let x^k be the sequence of iterates generated by the projected gradient descent of an L-smooth convex differentiable function f over a closed convex domain Ω. Then, using a fixed gradient step

$$\gamma = \frac{1}{L}$$

we have

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

Theorem 2.2

Let $f : \mathrm{dom}(f) \to \mathbb{R}$ be a convex differentiable L-smooth function over a closed and convex set $\Omega \subseteq \mathrm{dom}(f)$ and assume the existence of a minimizer $x^* \in \Omega$ of f in Ω. Given a constant stepsize

$$\gamma = \frac{1}{L},$$

the sequence of iterates of the projected gradient, starting at $x^0 \in \Omega$ satisfies

$$f(x^K) \leq f(x^*) + \frac{L}{2K}\|x^0 - x^*\|^2, \quad K > 0.$$

Exercise 2.2

Let $\Omega = B^{\ell_1}_{x^*}(R) := \{x \in \mathbb{R}^d : \|x - x^*\|_1 \leq R$ for some $R \leq 0$. Then $P_\Omega(v) = x^* + S_\theta(v - x^*)$, for all $v \in \mathbb{R}^d$ where

- θ is a parameter which will be defined in the proof
- $S_\theta(v)$ is the soft thresholding operator defined as

$$S_\theta(v)_i = \operatorname{sign}(v_i)(|v_i| - \theta)_+.$$

Remark 2.4

We are trying to solve the following problem:

$$P_{\ell_1,R}(u) = \operatorname{argmin} \|x-u\|_2 \quad \text{s.t. } \|x-x^*\|_1 \leq R$$

where x^* is a given centre and $R > 0$ a given radius.

Remark 2.5

Without loss of generality, we may assume that $x^* = 0$.

Remark 2.6

Without loss of generality, we may work with the following conditions:

1. $R = 1$,
2. $u_i \geq 0$, for all $1 \leq i \leq d$,
3. $\sum_{i=1}^{d} u_i > 1$.

Proposition 2.2

If $R = 1$ *and* $u_i \geq 0$ *for all* $1 \leq i \leq d$ *then* $y = P_{\ell_1}(u)$ *satisfies*

1. $y_i \geq 0$, *for all* $1 \leq i \leq d$ *and*
2. $\sum_{i=1}^{d} y_i = 1$.

Remark 2.7

Up to reshuffling of the indices, we may consider the entries of the vector u to be ordered:

$$u_1 \geq u_2 \geq \cdots \geq u_d.$$

Proposition 2.3

Let $u \in \mathbb{R}^d$ *and* $y = P_{\ell_1}(u)$, *with the remarks / assumptions from the previous results valid. Then there exists a unique* $p \in \{1, \cdots, d\}$ *such that*

- $y_i > 0$, *for* $1 \leq i \leq p$ *and*
- $y_i = 0$, *for* $p < i \leq d$.

Lemma 4

Let $u \in \mathbb{R}^d$ and define $y = P_{\ell_1}(u)$ with the conditions on u from the previous remarks. We have

$$y_i = u_i - \theta_p, \quad \text{for } 1 \leq i \leq p,$$

where

$$\theta_p = \frac{1}{p}\left(\sum_{i=1}^{p} -1\right).$$

Proposition 2.4

Let $u \in \mathbb{R}^d$ *be such that* $u_i \geq 0$*, for all* $1 \leq i \leq d$ *and* $\sum_{i=1}^d u_i > 1$*. For* $p \in \{1, \cdots, d\}$*, define* $y(p)$ *as*

$$y(p)_i = u_i - \theta_p, \quad 1 \leq i \leq p, \qquad \text{and } 0 \text{ for } i > p.$$

Then

$$y(p^*) = \operatorname{argmin}_{x \in \Delta_d} \|x - u\|_2$$

where

$$p^* = \max\{p : u_p - \frac{1}{p}\left(\sum_{i=1}^p u_p - 1\right) > 0\}$$

and Δ_d *defines the* d *dimensional unit simplex:*

$$\Delta_d = \left\{ x \in \mathbb{R}^d : x_i \geq 0 \text{ and } \sum_{i=1}^d x_i = 1 \right\}.$$

Theorem 2.3

The projection onto a ℓ_1 ball can be computed in $\mathcal{O}(d\log(d))$ operations.

Remark 2.8

The sorting problem can be reduced to a $\mathcal{O}(d)$. See John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra: *Efficient projections onto the ℓ_1-ball for learning in high dimensions*, in Proceedings of the 25th International Conference on Machine Learning, 2008.

Definition 2.2

The accelerated projected gradient descent is defined as

$$
\begin{aligned}
x^{k+1} &= P_\Omega(y^k - \frac{1}{L}\nabla f(y^k)) \\
z^{k+1} &= P_\Omega(z^k - \eta_k \nabla f(y^k)) \\
y^{k+1} &= (1 - \tau_{k+1})\, x^{k+1} + \tau_{k+1} z^{k+1},
\end{aligned}
$$

with $\tau_k = \frac{2}{k+2}$ *and* $\eta_k = \frac{k+1}{2L}$.

Theorem 2.4

Let Ω be a convex body and f a convex L-smooth function which admits a minimum $x^* \in \Omega$. The sequence of accelerated projected gradient descent starting at $x^0 = y^0 = z^0 \in \Omega$ satisfies

$$f(x^K) - f(x^*) \leq \frac{2L}{K(K+1)} \|x^0 - x^*\|_2^2, \quad \textit{for all } K \geq 1.$$