Projected Gradient Algorithm

Andersen Ang

ECS, Uni. Southampton, UK

andersen.ang@soton.ac.uk

Homepage angms.science

Version: April 11, 2024
First draft: August 2, 2017

Content

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} + \frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$$

$$f(\bar{\boldsymbol{x}}_K) - f^* \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{\sqrt{K+1}}.$$

Unconstrained minimization

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f(\boldsymbol{x}).$$

- All $\boldsymbol{x} \in \mathbb{R}^n$ is feasible.
- Any $\boldsymbol{x} \in \mathbb{R}^n$ can be a solution.

Constrained minimization

$$\min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}).$$

- Not all $\boldsymbol{x} \in \mathbb{R}^n$ is feasible.
- Not all $\boldsymbol{x} \in \mathbb{R}^n$ can be a solution.
- The solution has to be inside the set $\mathcal{Q}$.
- An example:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{x}\|_2 \leq 1$$

can be expressed as

$$\min_{\|\boldsymbol{x}\|_2 \leq 1} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2.$$

Here $\mathcal{Q} \coloneqq \{\boldsymbol{v} \in \mathbb{R}^n : \|\boldsymbol{v}\|_2 \leq 1\}$ is known as the unit ℓ_2 ball.

Approaches for solving constrained minimization problems

- Duality / Lagrangian approach
  - Not our focus here.
  - Although the approach of Lagrangian multiplier is usually taught in standard calculus class, the standard explanation that {gradient on primal variable has to be anti-parallel to gradient on the dual variable} is not intuitive and it is not the deep reason why the method works.
  - It requires a deep understanding of convex conjugate, constraint qualifications and duality to appreciate the Lagrangian approach, which is out of the scope here.

- First-order method / gradient-based method
  - **Simple**.
  - Our focus.

- Second-order method, Zero-order method, Higher-order method
  - Not our focus here.

Solving unconstrained problem $\min\limits_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$ by gradient descent

- Gradient descent **GD** is a $\begin{cases}\text{simple}\\ \text{easy}\\ \text{intuitive}\end{cases}$ way to solve **unconstrained** optimization problem $\min\limits_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$.

- Starting from an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$, **GD** iterates the following until a stopping condition is met:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k),$$

$k \in \mathbb{N}$: the current iteration counter
$k+1 \in \mathbb{N}$: the next iteration counter
$\boldsymbol{x}_k$: the current variable
$\boldsymbol{x}_{k+1}$: the next variable
∇f is the gradient of f with respect to differentiation of $\boldsymbol{x}$
$\nabla f(\boldsymbol{x}_k)$ is the ∇f at the current variable $\boldsymbol{x}_k$
$\alpha_k \in (0, +\infty)$: gradient stepsize

- **Question**: how about **constrained** problem? Is it possible to **tune GD** to fit constrained problem?
  Answer: yes, and the key is **Euclidean projection operator** $\mathrm{proj} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$.

- **Remark**
  - We assume f is differentiable (i.e., ∇f exists).
  - If f is not differentiable, we can replace gradient by subgradient, and we get the so-called subgradient method.

Problem setup of constrained problem

$$\min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}).$$

- We focus on the Euclidean space $\mathbb{R}^n$
- $f : \mathbb{R}^n \to \mathbb{R}$ is the objective / cost function
  - f is assumed to be continuously differentiable, i.e., $\nabla f(\boldsymbol{x})$ exists for all $\boldsymbol{x}$ $\quad f \in \mathcal{C}^1$
  - we assume f is globally L-Lipschitz, but not here $\quad |f(\boldsymbol{x}) - f(y)| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$
  - we do not assume ∇f is globally L-Lipschitz $\quad \|\nabla f(\boldsymbol{x}) - \nabla f(y)\| \leq L\|\boldsymbol{x} - \boldsymbol{y}\|$
- $\varnothing \neq \mathcal{Q} \subset \mathbb{R}^n$ is convex and compact
  - The constraint is represented by a set $\mathcal{Q}$
  - $\mathcal{Q} \subset \mathbb{R}^n$ means $\mathcal{Q}$ is a subset of $\mathbb{R}^n$, the domain of f
  - $\mathcal{Q} \neq \varnothing$ means $\mathcal{Q}$ is not an empty set $\quad$ it is not useful for discussion if $\mathcal{Q}$ is empty
  - $\mathcal{Q}$ is a convex set $\quad \forall \boldsymbol{x} \forall \boldsymbol{y} \forall \lambda \in (0,1) \Big\{ \boldsymbol{x} \in \mathcal{Q}, \boldsymbol{y} \in \mathcal{Q} \implies \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{Q} \Big\}$
  - $\mathcal{Q}$ is compact $\quad$ compact = bounded + closed
- For the details of convexity, Lipschitz, see here.

Solving constrained problem by projected gradient descent

- **Projected gradient descent PGD** $=$ **GD** $+$ projection
- Starting from an initial point $\boldsymbol{x}_0 \in \mathcal{Q}$, **PGD** iterates the following equation until a stopping condition is met:

$$\boldsymbol{x}_{k+1} = \mathcal{P}_{\mathcal{Q}}\Big(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)\Big),$$

  $k \in \mathbb{N}$: the current iteration counter
  $k+1 \in \mathbb{N}$: the next iteration counter
  $\boldsymbol{x}_k$: the current variable
  $\boldsymbol{x}_{k+1}$: the next variable
  ∇f is the gradient of f with respect to differentiation of $\boldsymbol{x}$
  $\nabla f(\boldsymbol{x}_k)$ is the ∇f at the current variable $\boldsymbol{x}_k$
  $\alpha_k \in (0, +\infty)$: gradient stepsize
  $\mathcal{P}_{\mathcal{Q}}$ is the shorthand of $\mathrm{proj}_{\mathcal{Q}}$

- $\mathrm{proj}_{\mathcal{Q}}(\,\cdot\,)$ is called **Euclidean projection operator**, and itself is also an optimization problem:

$$\mathcal{P}_{\mathcal{Q}}(\boldsymbol{x}_0) = \mathrm{proj}_{\mathcal{Q}}(\boldsymbol{x}_0) = \underset{\boldsymbol{x} \in \mathcal{Q}}{\operatorname{argmin}}\ \|\boldsymbol{x} - \boldsymbol{x}_0\|_2. \tag{$*$}$$

  i.e., given a point $\boldsymbol{x}_0$, $\mathcal{P}_{\mathcal{Q}}$ finds a point $\boldsymbol{x} \in \mathcal{Q}$ which is "closest" to $\boldsymbol{x}_0$.
  - The measure of "closeness" here is the Euclidean distance $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2$.

- $(*)$ is equivalent to

$$\underset{\boldsymbol{x} \in \mathcal{Q}}{\operatorname{argmin}}\ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2,$$

  where we squaring the cost so that the function becomes differentiable.

Comparing PGD to GD

GD

1. Pick an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$
2. Loop until stopping condition is met:
   - 2.1 Descent direction: compute $-\nabla f(\boldsymbol{x}_k)$
   - 2.2 Stepsize: pick a α_k
   - 2.3 Update: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$

PGD

1. Pick an initial point $\boldsymbol{x}_0 \in \mathcal{Q}$
2. Loop until stopping condition is met:
   - 2.1 Descent direction: compute $-\nabla f(\boldsymbol{x}_k)$
   - 2.2 Stepsize: pick a α_k
   - 2.3 Update: $\boldsymbol{y}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$
   - 2.4 Projection:
     $\boldsymbol{x}_{k+1} = \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{Q}} \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}_{k+1}\|_2^2$

- **PGD** $=$ **GD** $+$ projection.
  - if the point $\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$ after the gradient update is leaving the set $\mathcal{Q}$, project it back.
  - if the point $\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$ after the gradient update is within the set $\mathcal{Q}$, keep the point and do nothing.
- Projection $\mathcal{P}_{\mathcal{Q}}(\,\cdot\,) : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$
  - It is a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$, i.e., a point-to-point mapping
  - In general, for a nonconvex set $\mathcal{Q}$, such mapping is possibly non-unique (this is the $\rightrightarrows$)
  - $\mathcal{P}_{\mathcal{Q}}(\,\cdot\,)$ is an optimization problem

$$\mathcal{P}_{\mathcal{Q}}(\boldsymbol{x}_0) := \operatorname*{argmin}_{\boldsymbol{x} \in \mathcal{Q}} \frac{1}{2} \|\boldsymbol{x} - \boldsymbol{x}_0\|_2^2. \qquad (\star)$$

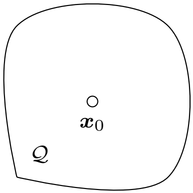
If $\mathcal{Q}$ is a convex compact set, the optimization problem has a unique solution, and we have $\mathcal{P}_{\mathcal{Q}}(\,\cdot\,) : \mathbb{R}^n \to \mathbb{R}^n$

- **PGD** is economic if $(\star)$ is $\begin{cases} \text{easy to solve} \\ \text{has a closed-form expression} \\ \text{cheap to compute} \end{cases}$
- **PGD** is **possibly not** an economic if $\begin{cases} \mathcal{Q} \text{ is nonconvex} \\ (*) \text{ has no closed-form expression} \\ (*) \text{ is expensive to compute} \end{cases}$

Understanding the geometry of projection ... (1/4)

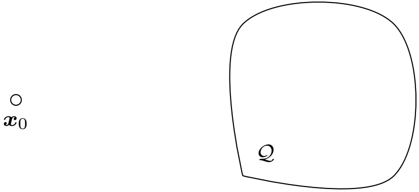
Consider a convex set $\mathcal{Q} \subset \mathbb{R}^n$ and a point $\boldsymbol{x}_0 \in \mathbb{R}^n$.

Case 1. $\boldsymbol{x}_0 \in \mathcal{Q}$.



- As $\boldsymbol{x}_0 \in \mathcal{Q}$, the closest point to $\boldsymbol{x}_0$ in $\mathcal{Q}$ will be $\boldsymbol{x}_0$ itself.
- The distance between a point to itself is zero.
- Mathematically: $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 = 0$ gives $\boldsymbol{x} = \boldsymbol{x}_0$.
- This is the trivial case and therefore not interesting.

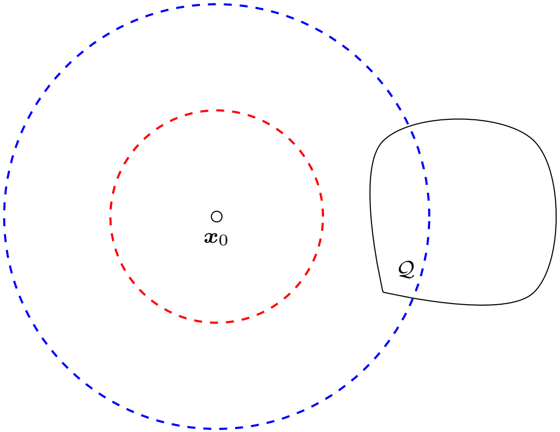
Case 2. $\boldsymbol{x}_0 \notin \mathcal{Q}$.



- Now $\boldsymbol{x}_0$ is outside $\mathcal{Q}$
- We need to find a point $\boldsymbol{x}$
  - $\boldsymbol{x} \in \mathcal{Q}$
  - $\|\boldsymbol{x} - \boldsymbol{x}_0\|_2$ is smallest
- This is case that is interesting.

Understanding the geometry of projection ... (2/4)

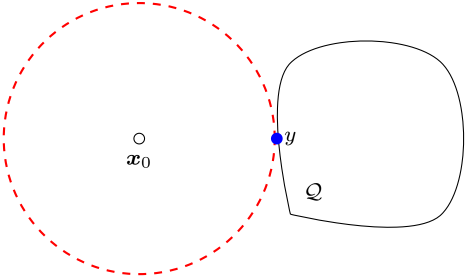
- The circles are ℓ_2-norm ball centered at $\boldsymbol{x}_0$ with different radius.
- Points on these circles are **equidistant** to $\boldsymbol{x}_0$ (with different l_2 distance on different circles).
- Note that some points on the blue circle are inside $\mathcal{Q}$, those are feasible points.



Understanding the geometry of projection ... (3/4)

- The point inside $\mathcal{Q}$ which is closest to $\boldsymbol{x}_0$ is the point where the ℓ_2 norm ball "touches" $\mathcal{Q}$.
- In this example, the blue point $\boldsymbol{y}$ is the solution to

$$\mathcal{P}_{\mathcal{Q}}(\boldsymbol{x}_0) = \operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2.$$

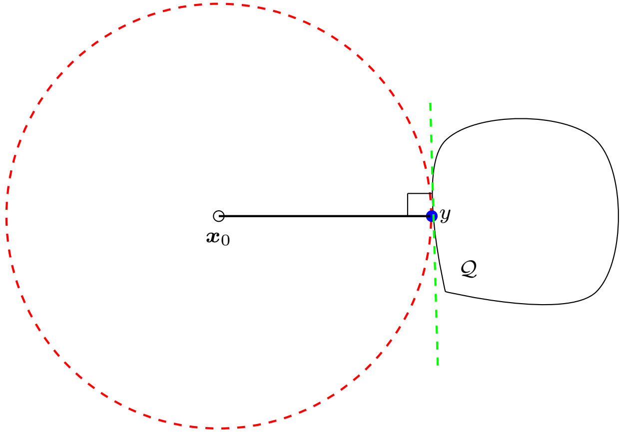


- In fact, such point is always located on the **boundary** of $\mathcal{Q}$ for $\boldsymbol{x}_0 \notin \mathcal{Q}$.
  That is, mathematically, if $\boldsymbol{x}_0 \notin \mathcal{Q}$, then

$$\operatorname*{argmin}_{\boldsymbol{x}\in\mathcal{Q}} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 \in \text{bdry}\mathcal{Q}.$$

Understanding the geometry of projection ... (4/4)

Note that the projection is **orthogonal**: the blue point $\boldsymbol{y}$ is always on a straight line that is tangent to the norm ball and $\mathcal{Q}$.



The normal to the tangent is exactly $\boldsymbol{x}_0 - \boldsymbol{y} = \boldsymbol{x}_0 - \mathrm{proj}_{\mathcal{Q}}(\boldsymbol{x}_0)$.

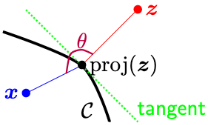
Property of projection: Bourbaki-Cheney-Goldstein inequality

Bourbaki-Cheney-Goldstein inequality[2]

- Modern names: Obtuse angle criterion, Projection theorem[1].
- What is it: a variational characterization of projection operator

$$\Big\langle \boldsymbol{z} - \operatorname{proj}(\boldsymbol{z}), \boldsymbol{x} - \operatorname{proj}(\boldsymbol{z}) \Big\rangle \leq 0, \quad \forall \boldsymbol{x} \in C.$$

The angle in between is obtuse ($\theta \geq 90°$).



[1]The name "projection theorem" is usually refer to projection in the context of Hilbert space (= vector space equipped with inner product, an operation that allows defining lengths and angles.)

[2]E. W. Cheney and A. A. Goldstein, Tchebycheff approximation and related extremal problems, J. Math. Mech. 14 (1965), 87-98.

4 / 9

Details here

PGD is a special case of proximal gradient

- The indicator function $\iota(\boldsymbol{x})$, of a set $\mathcal{Q}$ is defined as follows:

$$\iota_{\mathcal{Q}}(\boldsymbol{x}) = \begin{cases} 0 & \boldsymbol{x} \in \mathcal{Q} \\ +\infty & \boldsymbol{x} \notin \mathcal{Q} \end{cases}$$

- With the indicator function, constrained problem has two equivalent expressions

$$\min_{\boldsymbol{x} \in \mathcal{Q}} f(\boldsymbol{x}) \quad \equiv \quad \min_{\boldsymbol{x}} f(\boldsymbol{x}) + \iota_{\mathcal{Q}}(\boldsymbol{x}).$$

- Proximal gradient is a method to solve the optimization problem of a sum of differentiable and a non-differentiable function:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + g(\boldsymbol{x}),$$

  where g is non-differentiable.

- **PGD** is in fact the special case of proximal gradient where $g(\boldsymbol{x})$ is the indicator function. See here for more about proximal gradient .

On PGD ergodic convergence rate

- **Theorem 1**. If f is convex, PGD with constant stepsize α satisfies

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} + \frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$$

$\boldsymbol{x}^*$ is the (global) minimizer
$f^* := f(\boldsymbol{x}^*)$ is the optimal cost value
α is the constant stepsize
K is the total of number of iteration performed

- Interpretation:
  - the term $\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k$ is the "average" of the sequence $\boldsymbol{x}_k$ after K iterations
  - denote $\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k$ as $\bar{x}$
  - denote $f(\bar{x})$ as $\bar{f}$

  Then the theorem reads:

$$\bar{f} - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} + \text{something positive.}$$

  Hence the convergence rate is like $\mathcal{O}(\frac{1}{K})$.

- The term $\displaystyle\frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$ converges to zero
  - as long as $\displaystyle\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$ is not diverging to infinity, or
  - the growth of $\displaystyle\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$ is slower than K

What is ergodic convergence?

- Ergodic convergence = "The centroid of a point cloud moving towards the limit point"
- Sequence convergence: each of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ are all getting closer and closer to $\boldsymbol{x}^*$
- Ergodic convergence: the average of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ converges to $\boldsymbol{x}^*$
  - which doesn't imply each of $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_k$ are getting closer and closer to $\boldsymbol{x}^*$
  - some of them can be moving away from $\boldsymbol{x}^*$, as long as the centroid is getting closer

Proof of theorem 1 ... (1/3)

$$
\begin{array}{rll}
 & f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{x} \rangle & \mid f \text{ is convex} \\
\iff & f(\boldsymbol{x}) - f(\boldsymbol{z}) \leq \langle \nabla f(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{z} \rangle & \mid \\
\implies & f(\boldsymbol{x}_k) - f^* \leq \langle \nabla f(\boldsymbol{x}_k), \boldsymbol{x}_k - \boldsymbol{x}^* \rangle & \mid \boldsymbol{x} = \boldsymbol{x}_k, \boldsymbol{z} = \boldsymbol{x}^*, f(\boldsymbol{x}^*) = f^* \\
\iff & f(\boldsymbol{x}_k) - f^* \leq \left\langle \dfrac{\boldsymbol{x}_k - \boldsymbol{y}_{k+1}}{\alpha_k}, \boldsymbol{x}_k - \boldsymbol{x}^* \right\rangle & \mid \boldsymbol{y}_{k+1} \overset{\text{PGD}}{=} \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k) \\
\implies & f(\boldsymbol{x}_k) - f^* \leq \dfrac{\langle \boldsymbol{x}_k - \boldsymbol{y}_{k+1}, \boldsymbol{x}_k - \boldsymbol{x}^* \rangle}{\alpha} & \mid \text{constant stepsize}
\end{array}
$$

So we have

$$
\boxed{f(\boldsymbol{x}_k) - f^* \leq \frac{\langle \boldsymbol{x}_k - \boldsymbol{y}_{k+1}, \boldsymbol{x}_k - \boldsymbol{x}^* \rangle}{\alpha}}
$$

A not-so-trivial trick

$$
\begin{aligned}
(a-b)(a-c) &= a^2 - ac - ab + bc \\
&= \frac{2a^2 - 2ac - 2ab + 2bc}{2} \\
&= \frac{a^2 - 2ac + a^2 - 2ab + 2bc \color{red}{+ c^2 - c^2 + b^2 - b^2}}{2} \\
&= \frac{(a-c)^2 + (a-b)^2 - (b-c)^2}{2}
\end{aligned}
$$

Therefore

$$
\boxed{\langle \boldsymbol{x}_k - \boldsymbol{y}_{k+1}, \boldsymbol{x}_k - \boldsymbol{x}^* \rangle = \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{x}_k - \boldsymbol{y}_{k+1}\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2}}
$$

Combine the two boxes.

$$
f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + \|\boldsymbol{x}_k - \boldsymbol{y}_{k+1}\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha}
$$

$\boldsymbol{y}_{k+1} \overset{\text{PGD}}{=} \boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)$ we have $\boldsymbol{x}_k - \boldsymbol{y}_{k+1} = \alpha \nabla f(\boldsymbol{x}_k)$

Then

$$
f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 + \|\alpha \nabla f(\boldsymbol{x}_k)\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha}
$$

Now we have

$$
\boxed{f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2}
$$

Proof of theorem 1 ... (2/3)

Now we have

$$\boxed{f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2}$$

Next we need to make use of the fact that projection is non-expansive.

Explanation: focus on the term $\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2$

$\boldsymbol{x}_x$: current variable

$\boldsymbol{y}_{k+1}$: gradient updated $\boldsymbol{x}_k$

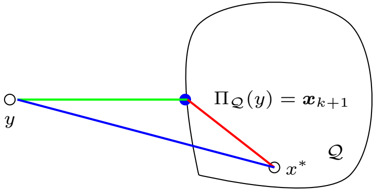
$\boldsymbol{x}_{k+1}$: projected $\boldsymbol{y}_{k+1}$

We wish to replace $\|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2$ by $\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2$
How: by the fact that projection operator is non-expansive

Note $\|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2 \geq \| \underbrace{\boldsymbol{x}_{k+1}}_{\text{proj}_{\mathcal{Q}}(\boldsymbol{y}_{k+1})} - \boldsymbol{x}^*\|_2^2$.

- This is known as "projection operator is non-expansive"
- "post-projection distance at most the same as the pre-projected"
- This is from the Bourbaki-Cheney-Goldstein inequality
- Details here

Pictorially



Hence $-\|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2 \leq -\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2$ and

$$\begin{aligned} f(\boldsymbol{x}_k) - f^* &\leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{y}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \\ &\leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2 \end{aligned}$$

It forms a telescoping series !

Proof of theorem 1 ... (3/3)

$$k = 0 \quad f(\boldsymbol{x}_0) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_1 - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_0)\|_2^2$$

$$k = 1 \quad f(x_1) - f^* \leq \frac{\|\boldsymbol{x}_1 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_2 - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_1)\|_2^2$$

$$\vdots$$

$$k = K \quad f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_k - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_{K+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\|\nabla f(\boldsymbol{x}_k)\|_2^2$$

Sums all

$$\sum_{k=0}^{K}\Big(f(\boldsymbol{x}_k) - f^*\Big) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

As $0 \leq \frac{1}{2\alpha}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}^*\|_2^2$,

$$\sum_{k=0}^{K}\Big(f(\boldsymbol{x}_k) - f^*\Big) \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha} + \frac{\alpha}{2}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Expand the summation on the left and divide the whole equation by $K + 1$

$$\frac{1}{K+1}\sum_{k=0}^{K} f(\boldsymbol{x}_k) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} + \frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2.$$

Consider the left hand side, as f is convex, by Jensen's inequality

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right) \leq \frac{1}{K+1}\sum_{k=0}^{K} f(\boldsymbol{x}_k).$$

Therefore

$$\begin{aligned} f\left(\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right) - f^* \leq\ & \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} \\ & + \frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2. \end{aligned}$$

$\square$

PGD converges ergodically at order $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$ on Lipschitz function

Theorem 2. If f is Lipschitz, for the point $\bar{\boldsymbol{x}}_K = \left\{ \frac{1}{K+1} \sum_{k=0}^{K} \boldsymbol{x}_k \right\}$ and constant stepsize $\alpha = \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{L\sqrt{K+1}}$ we have

$$f(\bar{\boldsymbol{x}}_K) - f^* \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{\sqrt{K+1}}.$$

Proof

- f is Lipschitz means ∇f is bounded: $\|\nabla f\| \leq L$, where L is the Lipschitz constant.
- Put $\bar{\boldsymbol{x}}_K$, α, $\|\nabla f\| \leq L$ into theorem 1.

Remarks

- On the stepsize α, note that it is K (total number of step) not k (current iteration number).
- α requires to know $\boldsymbol{x}^*$, so this theorem is practically useless as knowing $\boldsymbol{x}^*$ already solves the problem.
- Although we do not know $\boldsymbol{x}^*$ in general, the theorem tells that the ergodic convergence speed of **PGD** is $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

Discussion

In the convergence analysis of GD:

1. f is convex and β-smooth (gradient is β-Lipschitz)
2. Convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$.
3. The convergence rate is not ergodic

In the convergence analysis of PGD:

1. f is convex and L-Lipschitz (gradient is bounded above)
2. Convergence rate $\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$.
3. The convergence rate is ergodic, it works on $\bar{\boldsymbol{x}}_K$

If f is convex and β-smooth, the convergence of PGD will be the same as that of GD.

- Theoretical convergence rate of PGD on convex and β-smooth f is also $\mathcal{O}\left(\frac{1}{k}\right)$.
- However practically it depends on the complexity of the projection.
  Some $\mathcal{Q}$ are difficult to project onto.

As PGD is a special case of proximal gradient method, it is better to study proximal gradient method. For example here, here and here

Last page - summary

- PGD = GD + projection

- PGD with constant stepsize α:

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right) - f^* \leq \frac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2^2}{2\alpha(K+1)} + \frac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\boldsymbol{x}_k)\|_2^2$$

- IF f is Lipschitz (bounded gradient),

  for the point $\bar{\boldsymbol{x}}_K = \left\{\frac{1}{K+1}\sum_{k=0}^{K}\boldsymbol{x}_k\right\}$ and constant stepsize $\alpha = \dfrac{\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{L\sqrt{K+1}}$

  THEN

$$f(\bar{\boldsymbol{x}}_K) - f^* \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|}{\sqrt{K+1}}.$$

- **What's next: projection is possibly expensive, what about inexact projected gradient method?**

End of document