Projected Gradient Algorithm

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be   Homepage: angms.science

First draft: August 2, 2017
Last update : October 23, 2020

Overview

Constrained and unconstrained problem

- For unconstrained minimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}),$$

  any $\mathbf{x}$ in $\mathbb{R}^n$ can be a solution.

- For constrained minimization problem with a given set $\mathcal{Q} \subset \mathbb{R}^n$

$$\min_{\mathbf{x}\in\mathcal{Q}} f(\mathbf{x}),$$

  not any $\mathbf{x}$ can be a solution, the solution has to be inside the set $\mathcal{Q}$.

- An example of constrained minimization problem:

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

  can be expressed as

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

Solving unconstrained problem by gradient descent

- **Gradient Descent** (GD) is a standard (easy and simple) way to solve **unconstrained** optimization problem.
- Starting from an initial point $\mathbf{x}_0 \in \mathbb{R}^n$, GD iterates the following equation until a stopping condition is met:

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

  where ∇f is the gradient of f, the parameter $\alpha \geq 0$ is the step size, and k is the iteration counter.
- Question: how about **constrained** problem? Is it possible to **tune** GD to fit constrained problem?
  Answer: yes, and the key is **projection**.

Remark: If f is not differentiable, we can replace gradient by subgradient, and we get the so-called subgradient method.

Solving constrained problem by projected gradient descent

- **Projected Gradient Descent** (PGD) is a standard (easy and simple) way to solve **constrained** optimization problem.

- Consider a constraint set $\mathcal{Q} \subset \mathbb{R}^n$, starting from a initial point $\mathbf{x}_0 \in \mathcal{Q}$, PGD iterates the following equation until a stopping condition is met:

$$\mathbf{x}_{k+1} = P_{\mathcal{Q}}\Big(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)\Big).$$

- $P_{\mathcal{Q}}(\,.\,)$ is the projection operator, and itself is also an optimization problem:

$$P_{\mathcal{Q}}(\mathbf{x}_0) = \arg\min_{\mathbf{x}\in\mathcal{Q}} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2,$$

  i.e. given a point $\mathbf{x}_0$, $P_{\mathcal{Q}}$ try to find a point $\mathbf{x} \in \mathcal{Q}$ which is "closest" to $\mathbf{x}_0$.

About the projection

- $P_{\mathcal{Q}}(\,.\,)$ is a function from $\mathbb{R}^n$ to $\mathbb{R}^n$, and itself is an optimization problem:

$$P_{\mathcal{Q}}(\mathbf{x}_0) = \arg\min_{\mathbf{x}\in\mathcal{Q}} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2.$$

- PGD is an "economic" algorithm if the problem is easy to solve. This is not true for general $\mathcal{Q}$ and there are lots of constraint sets that are very difficult to project onto.

- If $\mathcal{Q}$ is a convex set, the optimization problem has a unique solution.

- If $\mathcal{Q}$ is nonconvex, the solution to $P_{\mathcal{Q}}(\mathbf{x}_0)$ may not be unique: it gives more than one solution.

Comparing PGD to GD

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

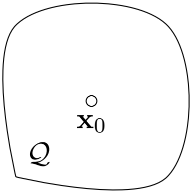
- PGD
  1. Pick an initial point $\mathbf{x}_0 \in \mathcal{Q}$
  2. Loop until stopping condition is met:
     - 2.1 Descent direction: pick the descent direction as $-\nabla f(\mathbf{x}_k)$
     - 2.2 Stepsize: pick a step size α_k
     - 2.3 Update: $\mathbf{y}_{k+1} = \mathbf{x}_k - \alpha_k \nabla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

- PGD has one more step: the projection.

- The idea of PGD is simple: if the point $\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ after the gradient update is leaving the set $\mathcal{Q}$, project it back.

Understanding the geometry of projection ... (1/5)

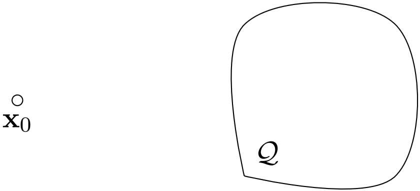
Consider a convex set $\mathcal{Q}$ and a point $\mathbf{x}_0 \in \mathcal{Q}$.



- As $\mathbf{x}_0 \in \mathcal{Q}$, the closest point to $\mathbf{x}_0$ in $\mathcal{Q}$ will be $\mathbf{x}_0$ itself.
- The distance between a point to itself is zero.
- Mathematically: $\frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 = 0$ gives $\mathbf{x} = \mathbf{x}_0$.

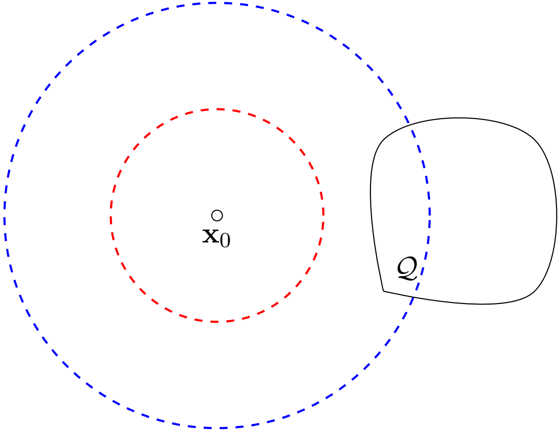
Understanding the geometry of projection ... (2/5)

Now consider a convex set $\mathcal{Q}$ and a point $\mathbf{x}_0 \notin \mathcal{Q}$: outside $\mathcal{Q}$.



Understanding the geometry of projection ... (3/5)

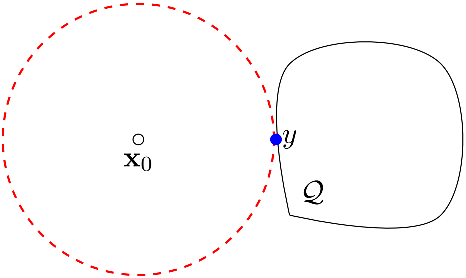
- The circles are L_2 norm ball centered at $\mathbf{x}_0$ with different radius.
- Points on these circles are **equidistant** to $\mathbf{x}_0$ (with different L_2 distance on different circles).
- Note that some points on the blue circle are inside $\mathcal{Q}$.



Understanding the geometry of projection ... (4/5)

- The point inside $\mathcal{Q}$ which is closest to $\mathbf{x}_0$ is the point where the L_2 norm ball "touches" $\mathcal{Q}$.
- In this example, the blue point $\mathbf{y}$ is the solution to

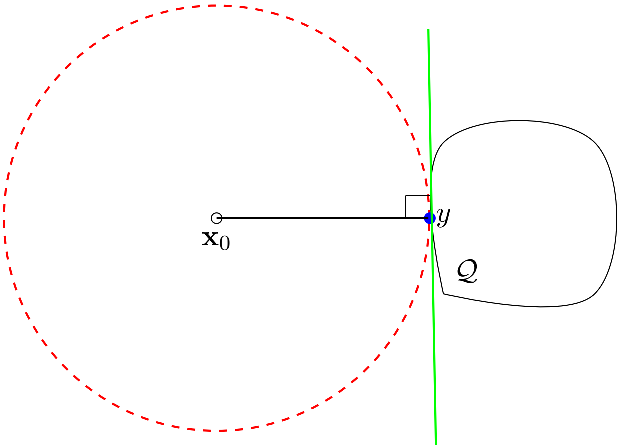
$$P_{\mathcal{Q}}(\mathbf{x}_0) = \underset{\mathbf{x}\in\mathcal{Q}}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2.$$



In fact, it can be proved that, such point is always located on the **boundary** of $\mathcal{Q}$ for $\mathbf{x}_0 \notin \mathcal{Q}$. That is, mathematically, $\underset{\mathbf{x}\in\mathcal{Q}}{\operatorname{argmin}} \ \frac{1}{2}\|\mathbf{x} - \mathbf{x}_0\|_2^2 \in \ \mathrm{bd}\mathcal{Q}$ if $\mathbf{x}_0 \notin \mathcal{Q}$.

Understanding the geometry of projection ... (5/5)

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



PGD is a special case of proximal gradient

- The indicator function, denoted as $i(\mathbf{x})$, of a set $\mathcal{Q}$ is defined as follows: if $\mathbf{x} \in \mathcal{Q}$, then $i(\mathbf{x}) = 0$; if $\mathbf{x} \notin \mathcal{Q}$, then $i(\mathbf{x}) = \infty$.
- With the indicator function, constrained problem has two equivalent expressions
$$\min_{\mathbf{x} \in \mathcal{Q}} f(\mathbf{x}) \quad \equiv \quad \min_{\mathbf{x}} f(\mathbf{x}) + i(\mathbf{x}).$$
- Proximal gradient is a method to solve the optimization problem of a sum of differentiable and a non-differentiable function:
$$\min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{x}),$$
where g is a non-differentiable function.
- PGD is in fact the special case of proximal gradient where $g(\mathbf{x})$ is the indicator function of the constrain set. See here for more about proximal gradient .

On PGD convergence rate

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

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

  where $f^* = f(\mathbf{x}^*)$ is the optimal cost value, $\mathbf{x}^*$ is the (global) minimizer, α is the constant stepsize, K is the total of number of iteration performed.
- Interpretation of this theorem: the term $\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{x}_k$ is the "average" of the sequence $\mathbf{x}_k$ after K iteration, hence we can denote it as $\bar{x}$ and $f(\bar{x})$ as $\bar{f}$. Then the theorem reads:

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

  Hence the convergence rate is like $\mathcal{O}(\frac{1}{K})$.
- For the second term on the right hand side, as long as $\sum_{k=0}^{K}\|\nabla f(\mathbf{x}_k)\|_2^2$ is not diverging to infinity, or the growth of it is slower than K, then the term $\dfrac{\alpha}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(\mathbf{x}_k)\|_2^2$ converges.

Proof of theorem 1 ... (1/5)

- f is convex so $f(\mathbf{x}) - f(\mathbf{z}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle$.

- Put $\mathbf{x} = \mathbf{x}_k$, $\mathbf{z} = \mathbf{x}^*$ and $f(\mathbf{x}^*) = f^*$:

$$f(\mathbf{x}_k) - f^* \leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - x^* \rangle.$$

- PGD update $\mathbf{y}_{k+1} = \mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k)$ gives $\nabla f(\mathbf{x}_k) = \dfrac{\mathbf{x}_k - \mathbf{y}_{k+1}}{\alpha_k}$ and

$$f(\mathbf{x}_k) - f^* \leq \frac{1}{\alpha_k} \langle \mathbf{x}_k - \mathbf{y}_{k+1}, \mathbf{x}_k - \mathbf{x}^* \rangle.$$

- As we use constant stepsize:

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

Proof of theorem 1 ... (2/5)

- A trick

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

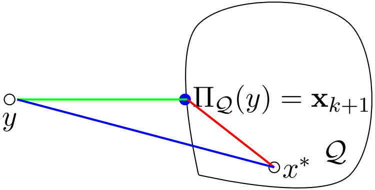
- Hence

$$
\begin{aligned}
f(\mathbf{x}_k) - f^* &\le \tfrac{1}{\alpha}\langle \mathbf{x}_k - \mathbf{y}_{k+1}, \mathbf{x}_k - \mathbf{x}^* \rangle \\
&= \tfrac{1}{2\alpha}\Big(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 + \|\mathbf{x}_k - \mathbf{y}_{k+1}\|_2^2 - \|\mathbf{y}_{k+1} - \mathbf{x}^*\|_2^2 \Big) \\
&\overset{*}{=} \tfrac{1}{2\alpha}\Big(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{y}_{k+1} - \mathbf{x}^*\|_2^2 \Big) + \tfrac{\alpha}{2}\|\nabla f(\mathbf{x}_k)\|_2^2
\end{aligned}
$$

where * is due to PGD update $\mathbf{x}_k - \mathbf{y}_{k+1} = \alpha \nabla f(\mathbf{x}_k)$

Proof of theorem 1 ... (3/5)

Note that $\|\mathbf{y}_{k+1} - \mathbf{x}^*\|_2^2 \geq \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2$.



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

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

It forms a telescoping series !

Proof of theorem 1 ... (4/5)

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

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

$$\vdots$$

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

Sums all

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

Proof of theorem 1 ... (5/5)

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

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

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

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

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

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

Therefore

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

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

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

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

Proof. Put $\bar{\mathbf{x}}_K$, α into theorem 1 directly, note that $\|\nabla f\| \leq L$.

Remarks

- f is Lipschitz then ∇f is bounded: $\|\nabla f\| \leq L$, where L is the Lipschitz constant.
- On the stepsize α, note that it is K (total number of step) not k (current iteration number).
- The stepsize requires to know $\mathbf{x}^*$, so this theorem is practically useless as knowing $\mathbf{x}^*$ already solves the problem.

Discussion

In the convergence analysis of GD:

1. f is convex and β-smooth (gradient is β-Lipschitz)
2. Convergence rate $\mathcal{O}\left(\frac{1}{k}\right)$.

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 works on $\bar{\mathbf{x}}_K$

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 will also be $\mathcal{O}\left(\frac{1}{k}\right)$.
- However practically it depends on the complexity of the projection. Some $\mathcal{Q}$ are difficult to project onto.

Last page - summary

- PGD = GD + projection

- PGD with constant stepsize α:

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

- If f is Lipschitz (bounded gradient), for the point $\bar{\mathbf{x}}_K = \left\{\frac{1}{K+1}\sum_{k=0}^{K}\mathbf{x}_k\right\}$ and constant step size $\alpha = \frac{\|\mathbf{x}_0 - \mathbf{x}^*\|}{L\sqrt{K+1}}$ then

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

End of document