Projected Gradient Algorithm

On convex function that is L-Lipschitz has convergence rate $\mathcal{Q}\left(\frac{1}{\sqrt{k}}\right)$

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft: August 2, 2017
Last update : July 29, 2019

Overview

Constrainted and unconstrained problem

Unconstrained minimization problem

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

Any $\mathbf{x}$ in the n-dimensional vector space $\mathbb{R}^n$ can be a solution.

Constrained minimization problem

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

x has to be inside the set $\mathcal{Q} \subset \mathbb{R}^n$.

Example of constrained minimization problem.

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \text{ such that } \|x\|_2 \leq 1$$

can be expressed as

$$\min_{\|x\|_2 \leq 1} \|Ax - b\|_2^2$$

Gradient Descent (GD) is a standard way to solve **unconstrained** optimization problem.
Starting from an initial point $x_0 \in \mathbb{R}^n$, GD iterates the following equation until a stopping condition is met :

$$x_{k+1} = x_k - t_k \nabla f(x_k)$$

Question : how about **constrained** problem ? Is it possible to **tune** GD to fit constrained problem ?

Answer : yes, and the key is **projection**.

Solving optimization problem by projected gradient descent

Projected Gradient Descent (PGD) is a way to solve **constrained** optimization problem. Consider a constraint set $\mathcal{Q}$, starting from a initial point $x_0 \in \mathcal{Q}$, PGD iterates the following equation until a stopping condition is met :

$$x_{k+1} = P_{\mathcal{Q}}\Big(x_k - t_k \nabla f(x_k)\Big)$$

where $P_{\mathcal{Q}}(\,.\,)$ is the projection operator

$$P_{\mathcal{Q}}(x_0) = \arg\min_{x \in \mathcal{Q}} \frac{1}{2}\|x - x_0\|_2^2$$

i.e. given a point x_0, $P_{\mathcal{Q}}$ try to find a point $x \in \mathcal{Q}$ which is "closest" to x_0.

Comparing PGD to GD

GD

1. Pick an inital point $x_0 \in \mathbb{R}^n$
2. Loop until stopping condition is met
   1. Descent direction : pick the descent direction as $-\nabla f(x_k)$
   2. Step size : pick a step size t_k
   3. Update : $x_{k+1} = x_k - t_k \nabla f(x_k)$

PGD

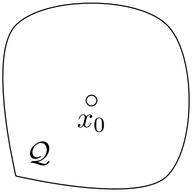
1. Pick an inital point $x_0 \in \mathcal{Q}$
2. Loop until stopping condition is met
   1. Descent direction : pick the descent direction as $-\nabla f(x_k)$
   2. Step size : pick a step size t_k
   3. Update : $y_{k+1} = x_k - t_k \nabla f(x_k)$
   4. Projection: $x_{k+1} = \arg\min_{x \in \mathcal{Q}} \frac{1}{2}\|x - y_{k+1}\|_2^2$

PGD has one more step: the projection.

The idea of PGD is simple: if the point $x_k - t_k \nabla f(x_k)$ after the gradient update is leaving the constraint set $\mathcal{Q}$, then project it back.

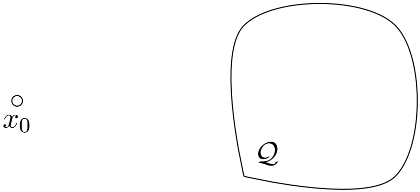
Understanding the geometry of projection - 1

Consider a convex set $\mathcal{Q}$ and a point $x_0 \in \mathcal{Q}$.



- As $x_0 \in \mathcal{Q}$, the closest point to x_0 in $\mathcal{Q}$ will be x_0 itself.
- The distance between a point and itself is zero.
- Mathematically, we have $\frac{1}{2}\|x - x_0\|_2^2 = 0$ which gives $x = x_0$.

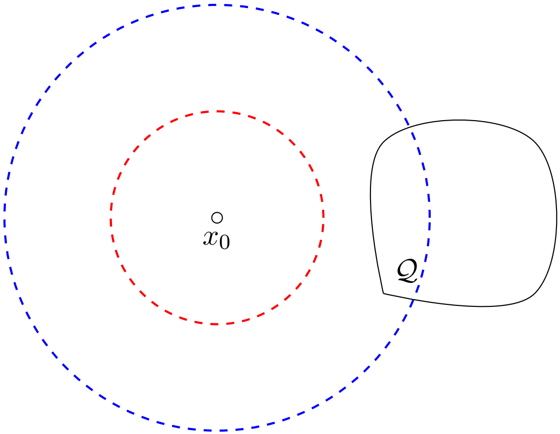
Now consider a convex set $\mathcal{Q}$ and a point $x_0 \notin \mathcal{Q}$.



As $\mathbf{x}_0 \notin \mathcal{Q}$, it is outside $\mathcal{Q}$.

Understanding the geometry of projection - 3

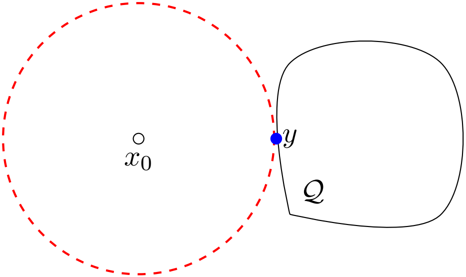
- The circles are L_2 norm ball centered at x_0 with different radius.
- Points on these circles are **equidistant** to x_0 (with different L_2 distance on different circles).
- Note that some points on the blue circle are inside $\mathcal{Q}$.



Understanding the geometry of projection - 4

- The point inside $\mathcal{Q}$ which is closest to x_0 is the point where the L_2 norm ball just "touch" $\mathcal{Q}$.
- In this example, the blue point y is the solution to

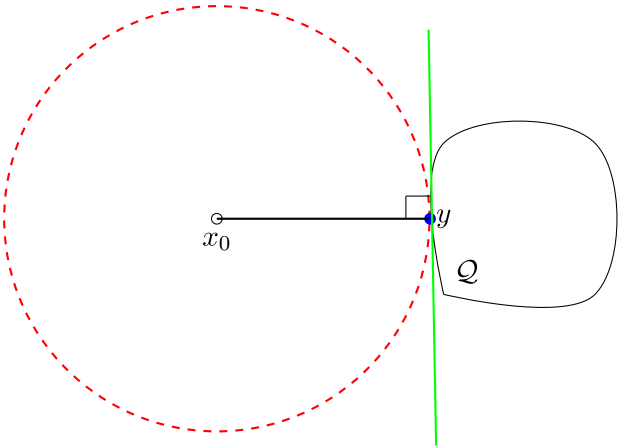
$$P_{\mathcal{Q}}(x_0) = \arg\min_{x \in \mathcal{Q}} \frac{1}{2}\|x - x_0\|_2^2$$



Actually it can be proved that such point is always located on the **boundary** of $\mathcal{Q}$ for x_0 outside $\mathcal{Q}$.

Understanding the geometry of projection - 5

Note that the point y is always on a straight line that is tangent to the norm ball and $\mathcal{Q}$.



On PGD convergence rate

Theorem 1. If f is convex, PGD with constant step size $t_k = t$ satisfies

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K} x_k\right) - f^* \le \frac{\|x_0 - x^*\|_2^2}{2t(K+1)} + \frac{t}{2(K+1)}\sum_{k=0}^{K} \|\nabla f(x_k)\|_2^2$$

Proof: f is convex $\iff f(y) \ge f(x) + \nabla f(x)^T (y - x)$, or

$$f(x) - f(y) \le \nabla f(x)^T (x - y)$$

Put $x = x_k$, $y = x^*$ and $f(x^*) = f^*$

$$f(x_k) - f^* \le \nabla f(x_k)^T (x_k - x^*)$$

By PGD update $y_{k+1} = x_k - t_k \nabla f(x_k)$ we get $\nabla f(x_k) = \dfrac{x_k - y_{k+1}}{t_k}$ and

$$f(x_k) - f^* \le \frac{1}{t_k}(x_k - y_{k+1})^T (x_k - x^*)$$

A trick

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

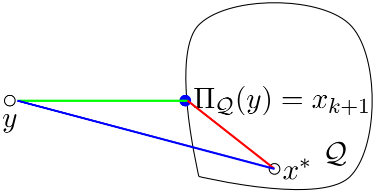
Hence

$$\begin{aligned}f(x_k) - f^* &\leq \frac{1}{t_k}(x_k - y_{k+1})^T(x_k - x^*) \\ &= \frac{1}{2t_k}\left(\|x_k - x^*\|_2^2 + \|x_k - y_{k+1}\|_2^2 - \|y_{k+1} - x^*\|_2^2\right)\end{aligned}$$

By PGD update again $x_k - y_{k+1} = t_k \nabla f(x_k)$ and thus

$$f(x_k) - f^* \leq \frac{1}{2t_k}\left(\|x_k - x^*\|_2^2 - \|y_{k+1} - x^*\|_2^2\right) + \frac{t_k}{2}\|\nabla f(x_k)\|_2^2$$

Note that $\|y_{k+1} - x^*\|_2^2 \geq \|x_{k+1} - x^*\|_2^2$.



Hence $-\|y_{k+1} - x^*\|_2^2 \leq -\|x_{k+1} - x^*\|_2^2$ and

$$\begin{aligned} f(x_k) - f^* &\leq \frac{1}{2t_k}\left(\|x_k - x^*\|_2^2 - \|y_{k+1} - x^*\|_2^2\right) + \frac{t_k}{2}\|\nabla f(x_k)\|_2^2 \\ &\leq \frac{1}{2t_k}\left(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\right) + \frac{t_k}{2}\|\nabla f(x_k)\|_2^2 \end{aligned}$$

It forms a telescoping series !

Proof of theorem 1 ... 4/5

$$k = 0 \qquad f(x_0) - f^* \leq \frac{\|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2}{2t_0} + \frac{t_0}{2}\|\nabla f(x_0)\|_2^2$$

$$k = 1 \qquad f(x_1) - f^* \leq \frac{\|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2}{2t_1} + \frac{t_1}{2}\|\nabla f(x_1)\|_2^2$$

$$\vdots$$

$$k = K \qquad f(x_K) - f^* \leq \frac{\|x_K - x^*\|_{K+1}^2 - \|x_2 - x^*\|_2^2}{2t_K} + \frac{t_K}{2}\|\nabla f(x_K)\|_2^2$$

Sums all, assuming constant step size $t_k = t$

$$\sum_{k=0}^{K}\Big(f(x_k) - f^*\Big) \leq \frac{\|x_0 - x^*\|_2^2 - \|x_{K+1} - x^*\|_2^2}{2t} + \frac{t}{2}\sum_{k=0}^{K}\|\nabla f(x_k)\|_2^2$$

Proof of theorem 1 ... 5/5

As $0 \leq \frac{1}{2t}\|x_{K+1} - x^*\|_2^2$

$$\sum_{k=0}^{K}\Big(f(x_k) - f^*\Big) \leq \frac{\|x_0 - x^*\|_2^2}{2t} + \frac{t}{2}\sum_{k=0}^{K}\|\nabla f(x_k)\|_2^2$$

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

$$\frac{1}{K+1}\sum_{k=0}^{K} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t(K+1)} + \frac{t}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(x_k)\|_2^2$$

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

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K} x_k\right) \leq \frac{1}{K+1}\sum_{k=0}^{K} f(x_k)$$

Therefore

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K} x_k\right) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t(K+1)} + \frac{t}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(x_k)\|_2^2 \quad \square$$

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

Theorem 2. If f is Lipschitz, for the point $\bar{x}_K \in \left\{ \dfrac{1}{K+1} \sum_{k=0}^{K} x_k \right\}$ and constant step size $t = \dfrac{\|x_0 - x^*\|}{L\sqrt{K+1}}$ we have

$$f(\bar{x}_K) - f^* \leq \frac{L\|x_0 - x^*\|}{\sqrt{K+1}}$$

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

Remarks

- The point $\bar{x}_K$ is the "average" of the x_k
- f is Lipschitz then ∇f is bounded: $\|\nabla f\| \leq L$, where L is the Lipschitz constant
- On the step size, note that it is K (total number of step) not k (current iteration number)
- The step size requires to know x^*, so this theorem is practically useless as knowing x^* already solves the optimization 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 convergen rate works on $\bar{x}_K$

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

1. PGD = GD + projection $P_{\mathcal{Q}}(x_0) = \arg\min_{x\in\mathcal{Q}} \frac{1}{2}\|x - x_0\|_2^2$
2. PGD with constant step size:

$$f\left(\frac{1}{K+1}\sum_{k=0}^{K} x_k\right) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t(K+1)} + \frac{t}{2(K+1)}\sum_{k=0}^{K}\|\nabla f(x_k)\|_2^2$$

3. If f is Lipschitz (bounded gradient), for the point $\bar{x}_K \in \left\{\frac{1}{K+1}\sum_{k=0}^{K} x_k\right\}$ and constant step size $t = \frac{\|x_0 - x^*\|}{L\sqrt{K+1}}$ then

$$f(\bar{x}_K) - f^* \leq \frac{L\|x_0 - x^*\|}{\sqrt{K+1}}$$

End of document