Proximal 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 16, 2019

Overview

1. Constrained problem, projected gradient and proximal gradient
2. Theorem. Proximal Gradient Descent converges at order $\mathcal{O}(\frac{1}{k})$
3. Summary

Constrained problem and projected gradient

Constrained minimization problem

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

Starting from an initial point $x_0 \in \mathcal{Q}$, Projected Gradient Descent (PGD) iterates

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

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

The proximal operator

The proximal operator $\text{prox}_{\alpha f}(x_0)$ of a function f, parameter α at the point x_0 is defined as

$$\text{prox}_{\alpha f}(x) = \arg\min_x \left\{ f(x) + \frac{1}{2\alpha}\|x - x_0\|_2^2 \right\}.$$

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

- Both are argmin function that returns a point x
- $P_{\mathcal{Q}}(x_0) = \text{prox}_f(x_0)$ with $f(x) = 0, \alpha = 1$
- For $P_{\mathcal{Q}}(x_0)$, x has to be inside $\mathcal{Q}$, for $\text{prox}_{\alpha f}(x)$, any x is feasible
- Projection operator : find a $x \in \mathcal{Q}$ that is closest to x_0.
- Proximal operator : find a x that "compromises" between the distance from x_0 and the minimization of f
- α controls the degree of compromises : larger α makes the second term less important
- For a sufficient large α, $f(x) + \frac{1}{2\alpha}\|x - x_0\|_2^2$ is convex even if f is non-convex

Comparing PGD and Proximal GD

Some constrained problems $\min_{x \in \mathcal{Q}} f(x)$ can be expressed as a unconstrained problem with augmented cost function as $\min_x f(x) + i_{\mathcal{Q}}(x)$, where $x \in \mathcal{Q}$ is expressed by the indicator function $i_{\mathcal{Q}}(x)$

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

To solve $\min_{x \in \mathcal{Q}} f(x)$, PGD step is

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

To solve $\min_x f(x) + i_{\mathcal{Q}}(x)$, the Proximal GD step is

$$x_{k+1} = \text{prox}_{\alpha i_{\mathcal{Q}}}\Big(x_k - t_k \nabla f(x_k)\Big)$$

$P_{\mathcal{Q}}(x)$ depends on $\mathcal{Q}$, $\text{prox}_{\alpha i_{\mathcal{Q}}}(x)$ depends on $i_{\mathcal{Q}}(x)$.
If $\mathcal{Q}$ and $i_{\mathcal{Q}}$ are complicated:

- hard to derive analytic expression of $P_{\mathcal{Q}}(x)$ and $\text{prox}_{\alpha i_{\mathcal{Q}}}(x)$.
- expensive to perform $P_{\mathcal{Q}}(x)$ and $\text{prox}_{\alpha g}(x)$

Normally $\mathcal{Q}$ and $i_{\mathcal{Q}}$ are assumed to be "simple" (e.g. convex, algebraic)

Example 1.

$$\min_x f(x) + \|x\|_2^2$$

The Proximal GD step is

$$\begin{aligned} x_{k+1} &= \text{prox}_{\alpha\|x\|_2^2}\Big(x_k - t_k \nabla f(x_k)\Big) \\ &= \arg\min_x \Big\{ \underbrace{\|x\|_2^2 + \frac{1}{2\alpha}\|x - x_k - t_k \nabla f(x_k)\|_2^2}_{F} \Big\} \end{aligned}$$

If f is differentiable, note that F is sum of two convex parts so it is convex.

To minimize F, set gradient $2x + \frac{1}{\alpha}\big(x - x_k - t_k \nabla f(x_k)\big)$ to 0 we get

$$x_{k+1} = \frac{1}{2\alpha + 1}\big(x_k - t_k \nabla f(x_k)\big)$$

Example 2.

$$\min_x f(x) + \|x\|_1$$

The Proximal GD update step is

$$\begin{aligned} x_{k+1} &= \text{prox}_{\alpha\|x\|_1}\Big(x_k - t_k \nabla f(x_k)\Big) \\ &= \arg\min_x \left\{ \|x\|_1 + \frac{1}{2\alpha}\|x - x_k - t_k \nabla f(x_k)\|_2^2 \right\} \\ &= \arg\min_x \left\{ \sum_i^n \left(|x_i| + \frac{1}{2\alpha}(x_i - [x_k]_i - t_k[\nabla f(x_k)]_i)^2 \right) \right\} \end{aligned}$$

which becomes the element-wise soft-thresholding

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

where

$$[S_\alpha(x)]_i = \begin{cases} x_i - \alpha & x_i \geq \alpha \\ 0 & |x_i| \leq \alpha \\ x_i + \alpha & x_i \leq -\alpha \end{cases}$$

Comparing PGD and Proximal GD - on algorithm

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

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: $x_{k+1} = \arg\min_{x \in \mathcal{Q}} \frac{1}{2} \|x - y_{k+1}\|_2^2$

Proximal GD (for $\min_x f(x) + g(x)$)

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 : $y_{k+1} = x_k - t_k \nabla f(x_k)$
   4. Pick the α parameter
   5. Solving optimization : $x_{k+1} = \arg\min_x \left\{ g(x) + \frac{1}{2\alpha} \|x - y_{k+1}\|_2^2 \right\}$

Problem setting for Proximal GD

Generally

$$\min_x F(x) = f(x) + g(x)$$

where

- f is convex, differentiable (possibly β-smooth, α-strongly convex)
- g is convex (possibly non-differentiable), $\text{prox}_{\alpha g}(x)$ is inexpensive
- F^* is finite and exists (at minimizer x^*, but may not be unique)

Then for the update

$$x_{k+1} = \text{prox}_{\alpha g}\Big(x_k - t_k \nabla f(x_k)\Big)$$

- the solution to $\text{prox}_{\alpha g}\big(x_k - t_k \nabla f(x_k)\big)$ exists and unique, as g is convex (and close)
- i.e., the update provide an unique solution x_{k+1} every time

Gradient mapping of Proximal GD update

$$
\begin{aligned}
x_{k+1} &= \mathrm{prox}_{\alpha g}\Big(x_k - t_k \nabla f(x_k)\Big) \\
&\overset{\text{tricky}}{=} x_k - \Big\{x_k - \mathrm{prox}_{\alpha g}\Big(x_k - t_k \nabla f(x_k)\Big)\Big\} \\
&= x_k - t_k \frac{1}{t_k}\Big\{x_k - \mathrm{prox}_{\alpha g}\Big(x_k - t_k \nabla f(x_k)\Big)\Big\} \\
&= x_k - t_k G_{t_k}(x_k)
\end{aligned}
$$

where

$$
G_t(x) = \frac{1}{t}\Big[x - \mathrm{prox}_{tg}\Big(x - t\nabla f(x)\Big)\Big].
$$

- The expression $x_{k+1} = t_k G_{t_k}(x_k)$ is called the gradient mapping of Proximal GD.

1 page review on sub-gradient

For a differentiable convex f, for any x, y we have

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y - x)}_{\phi(x)},$$

the expression $\phi(x)$, which is the first order Taylor approximation, forms a global underestimator of f at the point x for all y.

Similarly, for a non-differentiable convex h, for any x, y we have

$$h(y) \geq h(x) + g^T (y - x),$$

the vector g "act as" ∇f is called the *subgradient*

- subgradient always exists
- unlike gradient, function can have multiple subgradients at a point
- the set of all subgradients at x is the *sub-differential* $\partial h(x)$
- $q \in \partial h(x) \iff h(y) \geq h(x) + q^T (y - x)$.
  In words : given a function h, a vector q is a sub-gradient of h at the point x if $h(y) \geq h(x) + q^T (y - x)$ holds $\forall y$

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (1/6)

Theorem. For $\min_x F(x) = f(x) + g(x)$ where f is β-smooth and convex, g is convex (possibly non-differentiable), the sequence $F(x_k)$ produce by proximal GD converges to $F^* = F(x^*)$ at rate $\mathcal{O}(\frac{1}{k})$:

$$F(x_k) - F^* \leq \frac{\beta \|x_0 - x^*\|_2^2}{2k}$$

Proof: Consider f

$$f \text{ is } \beta\text{-smooth} \qquad f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2}\|x - y\|_2^2 \tag{1}$$

$$f \text{ is convex} \qquad f(x) \leq f(z) - \nabla f(x)^T (z - x) \tag{2}$$

$$(1) + (2) \qquad f(y) \leq f(z) - \nabla f(x)^T (z - y) + \frac{\beta}{2}\|x - y\|_2^2 \tag{3}$$

Put the gradient mapping $y = x - tG_t(x)$ into (3)

$$f\big(x - tG_t(x)\big) \quad \leq \quad f(z) - \nabla f(x)^T \big(z - x + tG_t(x)\big) + \frac{\beta t^2}{2}\|G_t(x)\|_2^2 \tag{4}$$

As $F(x) = f(x) + g(x)$, by (4)

$$F\big(x - tG_t(x)\big) \leq f(z) - \nabla f(x)^T \big(z - x + tG_t(x)\big) + \frac{\beta t^2}{2}\|G_t(x)\|_2^2 + g\big(x - tG_t(x)\big)$$

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (2/6)

To handle $g\big(x - tG_t(x)\big)$, we show

$$g\big(x - tG_t(x)\big) \leq g(z) - (G_t(x) - \nabla f(x))^T \Big(z - x + tG_t(x))\Big)$$

The key to handle $g\big(x - tG_t(x)\big)$ is to use subgradient : note that g is non-differentiable, but it is convex, so subgradient inequality holds

$$g(z) \quad \geq \quad g(x - tG_t(x)) + q^T \Big(z - (x - tG_t(x))\Big)$$

In words: there is an affine global underestimator formed by q (the subgradient) at point $x - tG_t(x)$ for all z.

Re-arrange the inequality we get

$$g(x - tG_t(x)) \leq g(z) - q^T \Big(z - x + tG_t(x))\Big) \tag{5}$$

Now the question is : how to find the sub-gradient q?

Showing $g(x - tG_t(x)) \leq g(z) - (G_t(x) - \nabla f(x))^T \Big(z - x + tG_t(x))\Big)$ ... (1/2)

Answer: find q by optimality condition.
Recall the equivalence between gradient mapping and the proximal operator

$$\begin{aligned} y &= x - tG_t(x) = \text{prox}_{tg}\Big(x - t\nabla f(x)\Big) \\ &= \arg\min_w g(w) + \frac{1}{2t}\|w - x + t\nabla f(x)\|_2^2 \end{aligned} \tag{6}$$

The sub-gradient optimality condition of (6) is

$$0 \in \partial g(w') + \partial \frac{1}{2t}\|w' - x + t\nabla f(x)\|_2^2.$$

Once again, the reason why sub-gradient is used here is because g is (possibly) not differentiable, however $\frac{1}{2t}\|w' - x + t\nabla f(x)\|_2^2$ is differentiable and the sub-gradient can be replaced by the normal gradient :

$$\begin{aligned} 0 &\in \partial g(w') + \nabla_{w'} \frac{1}{2t}\|w' - x + t\nabla f(x)\|_2^2 \\ &= \partial g(w') + \frac{1}{t}(w' - x + t\nabla f(x)). \end{aligned} \tag{7}$$

Showing $g(x - tG_t(x)) \leq g(z) - (G_t(x) - \nabla f(x))^T \Big(z - x + tG_t(x))\Big)$ ... (2/2)

Re-arrange (7) gives

$$\frac{1}{t}(x - w') - \nabla f(x) \in \partial g(w'). \tag{8}$$

As w' is the updated point : $w' = \text{prox}_{tg}\Big(x - t\nabla f(x)\Big)$ so

$$\frac{1}{t}(x - w') = \frac{1}{t}\Big[x - \text{prox}_{tg}\Big(x - t\nabla f(x)\Big)\Big] = G_t(x) \tag{9}$$

Put (9) into (8) gives

$$G_t(x) - \nabla f(x) \in \partial g(w').$$

In words, $G_t(x) - \nabla f(x)$ is one of the sub-gradient of g at the point w' (which is $y = x - tG_t(x)$).

Therefore, we have $q = G_t(x) - \nabla f(x)$ in (5)

$$g(x - tG_t(x)) \leq g(z) - (G_t(x) - \nabla f(x))^T \Big(z - x + tG_t(x))\Big) \tag{10}$$

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (3/6)

Previously we have

$$F\big(x - tG_t(x)\big) \leq f(z) - \nabla f(x)^T\big(z - x + tG_t(x)\big) + \frac{\beta t^2}{2}\|G_t(x)\|_2^2 + g\big(x - tG_t(x)\big)$$

and by (10)

$$g(x - tG_t(x)) \leq g(z) - (G_t(x) - \nabla f(x))^T\Big(z - x + tG_t(x))\Big)$$

Combine the two

$$\begin{aligned}
F\Big(x - tG_t(x)\Big) &\leq f(z) + g(z) - G_t(x)^T\Big(z - x + tG_t(x)\Big) + \frac{\beta t^2}{2}\|G_t(x)\|_2^2 \\
\begin{pmatrix} F = f + g \\ \text{put } z = x^* \end{pmatrix} &= F^* - G_t(x)^T\Big(x^* - x + tG_t(x)\Big) + \frac{\beta t^2}{2}\|G_t(x)\|_2^2 \\
(\text{let } t \leq \frac{1}{\beta}) &\leq F^* - G_t(x)^T\Big(x^* - x + tG_t(x)\Big) + \frac{t}{2}\|G_t(x)\|_2^2 \\
(\text{combine } G^TG) &= F^* - G_t(x)^T(x^* - x) - \frac{t}{2}\|G_t(x)\|_2^2
\end{aligned}$$

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (4/6)

The most tricky step, pure algebra trick

$$-G_t(x)^T(x^* - x) - \frac{t}{2}\|G_t(x)\|_2^2 = \frac{1}{2t}\Big(\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\Big)$$

(note that $x^* - x$ becomes $x - x^*$ in the second term on right hand side)

Hence

$$\begin{aligned} F\Big(x - tG_t(x)\Big) &\leq F^* - G_t(x)^T(x^* - x) - \frac{t}{2}\|G_t(x)\|_2^2 \\ &\leq F^* + \frac{1}{2t}\Big(\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\Big) \end{aligned}$$

Finally, put $x = x_k$ and recall $x_k - t_k G_{t_k}(x_k) = x_{k+1}$

$$F(x_{k+1}) - F^* \leq \frac{1}{2t_k}\Big(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2\Big)$$

It forms a telescoping series !

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (5/6)

$$
\begin{aligned}
k = 0 \qquad & F(x_1) - F^* \leq \frac{1}{2t_0}\Big(\|x_0 - x^*\|_2^2 - \|x_1 - x^*\|_2^2\Big) \\
k = 1 \qquad & F(x_2) - F^* \leq \frac{1}{2t_1}\Big(\|x_1 - x^*\|_2^2 - \|x_2 - x^*\|_2^2\Big) \\
\vdots \qquad & \qquad\qquad\qquad \vdots \\
k = K-1 \qquad & F(x_K) - F^* \leq \frac{1}{2t_{K-1}}\Big(\|x_{K-1} - x^*\|_2^2 - \|x_K - x^*\|_2^2\Big)
\end{aligned}
$$

Let step size be constant $t_k = t \leq \frac{1}{\beta}$, sum all the equations gives

$$
\sum_{i=0}^{K-1}\Big(F(x_i) - F^*\Big) \leq \frac{1}{2t}\Big(\|x_0 - x^*\|_2^2 - \|x_K - x^*\|_2^2\Big)
$$

As $0 \leq \frac{1}{2}\|x_K - x^*\|_2^2$, add this into the previous inequality we get

$$
\sum_{i=0}^{K-1}\Big(F(x_i) - F^*\Big) \leq \frac{1}{2t}\|x_0 - x^*\|_2^2.
$$

Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ ... (6/6)

By definition, the method produces a non-increasing sequence $F(x_i)$

$$F(x_K) \leq ... \leq F(x_i) \leq ... \leq F(x_2) \leq F(x_1) \leq F(x_0)$$

Hence

$$\sum_{i=0}^{K-1} \Big(F(x_K) - F^* \Big) \leq \sum_{i=0}^{K-1} \Big(F(x_i) - F^* \Big) \leq \frac{1}{2t} \|x_0 - x^*\|_2^2$$

Ignore the middle term

$$\sum_{i=0}^{K-1} \Big(F(x_K) - F^* \Big) \leq \frac{1}{2t} \|x_0 - x^*\|_2^2$$

Note $F(x_K)$ and F^* are independent of the subscript i,

$$K\Big(F(x_K) - F^* \Big) \leq \frac{1}{2t} \|x_0 - x^*\|_2^2$$

Re-arrange and put $t = \dfrac{1}{\beta}$

$$F(x_K) - F^* \leq \frac{\beta \|x_0 - x^*\|_2^2}{2K}. \quad \square$$

Last page - summary

1. Proximal operator of a function f, parameter α and a given point x_0:

$$\text{prox}_{\alpha f}(x) = \arg\min_x \left\{ f(x) + \frac{1}{2\alpha}\|x - x_0\|_2^2 \right\}$$

2. Proximal GD solves $\min_x F(x) = f(x) + g(x)$ where f, g are convex but f is differentiable g is not.

3. Proximal GD update formula $x_{k+1} = \text{prox}_{\alpha g}(x_k - t_k \nabla f(x_k))$

4. Gradient mapping $G_t(x) = \frac{1}{t}\Big[x - \text{prox}_{\alpha g}(x - t\nabla f(x))\Big]$

5. Proximal GD converges at order $\mathcal{O}(\frac{1}{k})$ with constant step size $t_k = \frac{1}{\beta}$:

$$F(x_k) - F^* \leq \frac{\beta\|x_0 - x^*\|_2^2}{2k}.$$

End of document