# (Fast) dual proximal gradient algorithm

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft : March 20, 2020
Last update : March 21, 2020

# Problem setting

$$(P) \quad \min_{\mathbf{x}} f(\mathbf{x}) + g(\mathcal{A}\mathbf{x})$$

- $f : \mathbb{E} \to \bar{\mathbb{R}}$, $g : \mathbb{V} \to \bar{\mathbb{R}}$
- $\mathrm{dom} f = \mathbb{E}$ and $\mathrm{dom} g = \mathbb{V}$. $\mathbb{E}, \mathbb{V}$ are Euclidean spaces with inner products and norms[1]
- $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ is the extended real line
- $f$ is proper, close and strongly convex
- $g$ is proper, close and convex, possibly nonsmooth
- $\mathcal{A} : \mathbb{E} \to \mathbb{V}$ is a linear operator.
- By properties of $f$ and $g$, problem $(P)$ has a unique optimal sol. denoted by $\mathbf{x}^*$.

[1] For simplicity, can treat $\mathbb{E}, \mathbb{V}$ as some subset of $\mathbb{R}^n$.

# Example : denoising

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{d}\|_2^2 + \lambda R(\mathcal{L}\mathbf{x})$$

- $\mathbf{d} \in \mathbb{R}^n$ is a given signal.
- $R$ is a convex function.
- $\mathcal{L}$ is a linear transformation accounts for smoothness of signal.
- An example : total-variation norm, where $R(\mathcal{L}\mathbf{x}) = \|\mathbf{D}\mathbf{x}\|_1$, with $\mathbf{D}$ as the difference operator and $R$ is the $L_1$ norm.
- Note that there is no simple projection for the constraint type $\|\mathbf{D}\mathbf{x}\|_1 \leq \epsilon$. The proximal operator involving $\|\mathbf{D}\mathbf{x}\|_1$ has no simple analytic close form solution. These motivates the use of dual problem.
- See here for details on solving the total variation norm regularized least square problem using majorization-minimization.

# Example : projection onto intersection of convex sets

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{d}\|_2^2 \text{ s.t. } \mathbf{x} \in \bigcap_{i=1}^{m} C_i$$

- $\mathbf{d} \in \mathbb{R}^n$ is a given, and $C_1, C_2, \cdots, C_m$ are convex sets with nonempty intersection
- The problem ask to find the closest point of $\mathbf{d}$ inside the intersection of $C_i$.
- Using indicator function $i_{C_i}(\cdot)$ for each set $C_i$, we can express this problem in the form of $(P)$ as

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{d}\|_2^2 + \sum_{i=1}^{m} i_{C_i}(\mathbf{x})$$

where $f(\mathbf{x}) = \|\mathbf{x}-\mathbf{d}\|_2^2$, $g(\mathcal{A}\mathbf{x}) = g(\mathbf{x}, \mathbf{x}, \cdots, \mathbf{x}) = \sum_{i=1}^{m} i_{C_i}(\mathbf{x})$ with $\mathcal{A} : \mathbb{E} \to \mathbb{E}^m$ defined as $\mathcal{A}(\mathbf{x}) = \underbrace{(\mathbf{x}, \mathbf{x}, \cdots, \mathbf{x})}_{m \text{ blocks}}$

- Note that the proximal operator involving $\sum_{i=1}^{m} i_{C_i}(\mathbf{x})$ has no simple analytic close form solution. These motivates the use of dual problem.

# Example : projection onto polyhedron

$$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{d}\|_2^2 \text{ s.t. } \mathbf{A}\mathbf{x} \leq \mathbf{b}$$

- projection onto polyhedron is a special case of projection onto intersection of convex sets, where the convex sets $\mathbf{C}_i$ are now halfspaces $\langle \mathbf{a}^i, \mathbf{x} \rangle \leq b_i$, with $\mathbf{a}^i$ denoting the $i^{\text{th}}$ row of $\mathbf{A} \in \mathbb{R}^{m \times n}$, and $b_i$ is the $i^{\text{th}}$ element of $\mathbf{b} \in \mathbb{R}^m$.
- Using indicator function $i_{(-\infty, b_i]}(\mathbf{y})$ for each set $\{\mathbf{y} \mid \langle \mathbf{a}^i, \mathbf{x} \rangle \leq b_i\}$, we can express this problem in the form of $(P)$ as

  $$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{d}\|_2^2 + \sum_{i=1}^{m} i_{(-\infty, b_i]}(\mathbf{x})$$

  where $\mathcal{A} : \mathbb{E} \to \mathbb{E}^m$ defined as $\mathcal{A}(\mathbf{x}) = \underbrace{(\mathbf{x}, \mathbf{x}, \cdots, \mathbf{x})}_{m \text{ blocks}}$ Or, using indicator function $i_{(-\infty, \mathbf{b}]}(\mathbf{y})$ for the set $\{\mathbf{y} \mid \mathbf{y} \leq \mathbf{b}\}$, we can express this problem in the form of $(P)$ as

  $$\min_{\mathbf{x}} \|\mathbf{x}-\mathbf{d}\|_2^2 + i_{(-\infty, \mathbf{b}]}(\mathbf{A}\mathbf{x})$$

  where $\mathcal{A}(\mathbf{x}) = \mathbf{A}\mathbf{x}$.
- As discussed here, this problem has no simple close form sol. This motives the use of dual problem.

# Deriving the dual formulation of $(P)$ ... 1/2

- Problem $(P)$ written in constrained form

$$(P') \quad \min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{z}) \text{ s.t. } \mathcal{A}\mathbf{x} = \mathbf{z}.$$

- The Lagrangian

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} - \mathbf{z} \rangle,$$

  $\mathbf{y}$ is the Lagrangian multiplier associated to the constraint in $(P')$.
- Recall that minimizing the Lagrangian w.r.t. $\mathbf{x}$ and $\mathbf{z}$ gives the dual problem. Therefore it is better to split the Lagrangian into two parts, one part focus on $\mathbf{x}$ and the other part focus on $\mathbf{z}$

$$\begin{aligned} L(\mathbf{x}, \mathbf{z}, \mathbf{y}) &= f(\mathbf{x}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} \rangle + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} \rangle. \\ \min_{\mathbf{x},\mathbf{z}} L &= \min_{\mathbf{x},\mathbf{z}} \Big\{ f(\mathbf{x}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} \rangle + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} \rangle \Big\} \\ &= \min_{\mathbf{x}} \Big\{ f(\mathbf{x}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} \rangle \Big\} + \min_{\mathbf{z}} \Big\{ g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} \rangle \Big\} \end{aligned}$$

- Next we use convex conjugate to simplify the notation

# Deriving the dual formulation of $(P)$ ... 2/2

- Recall $\min f(\mathbf{x}) = -\max\{-f(\mathbf{x})\}$

$$\begin{aligned}
\min_{\mathbf{x},\mathbf{z}} L &= \min_{\mathbf{x}} \Big\{ f(\mathbf{x}) - \langle \mathbf{y}, \mathcal{A}\mathbf{x} \rangle \Big\} + \min_{\mathbf{z}} \Big\{ g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{z} \rangle \Big\} \\
&= -\max_{\mathbf{x}} \Big\{ -f(\mathbf{x}) + \langle \mathbf{y}, \mathcal{A}\mathbf{x} \rangle \Big\} - \max_{\mathbf{z}} \Big\{ -g(\mathbf{z}) - \langle \mathbf{y}, \mathbf{z} \rangle \Big\} \\
&= -\max_{\mathbf{x}} \Big\{ -f(\mathbf{x}) + \langle \mathcal{A}^\top \mathbf{y}, \mathbf{x} \rangle \Big\} - \max_{\mathbf{z}} \Big\{ -g(\mathbf{z}) + \langle -\mathbf{y}, \mathbf{z} \rangle \Big\}
\end{aligned}$$

- For a function $h(\mathbf{u})$, the convex conjugate $h^*(\mathbf{u})$ is defined as

$$h^*(\mathbf{u}) = \max_{\mathbf{v}} \big\{ \langle \mathbf{u}, \mathbf{v} \rangle - f(\mathbf{v}) \big\}.$$

  Let $f^*$ and $g^*$ be the conjugate of $f$ and $g$ respectively, we have

$$\min_{\mathbf{x},\mathbf{z}} L = -f^*(\mathcal{A}^\top \mathbf{y}) - g^*(-\mathbf{y})$$

- The dual problem is thus

$$(D) \quad \max_{\mathbf{y}} q(\mathbf{y}) = -f^*(\mathcal{A}^\top \mathbf{y}) - g^*(-\mathbf{y}).$$

  In minimization form

$$(D') \quad \min_{\mathbf{y}} f^*(\mathcal{A}^\top \mathbf{y}) + g^*(-\mathbf{y}).$$

# Duality and $f^*$ has Lipschitz gradient

- Recall Slater condition : if there exists $\mathbf{x} \in \text{ri}(\text{dom } f)$ and $\mathbf{z} \in \text{ri}(\text{dom } g)$ s.t. $\mathcal{A}\mathbf{x} = \mathbf{z}$, then strong duality holds, and optimal sol. of dual problem is attained.
- Recall $f$ is strongly convex with modulus $\sigma$. This implies $f^*$ has Lipschitz gradient : let $F(\mathbf{y}) = f^*(\mathcal{A}^\top \mathbf{y})$,

$$\begin{aligned}
\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| &= \|\mathcal{A}\nabla f^*(\mathcal{A}^\top \mathbf{x}) - \mathcal{A}\nabla f^*(\mathcal{A}^\top \mathbf{y})\| \\
&\leq \|\mathcal{A}\| \cdot \|\nabla f^*(\mathcal{A}^\top \mathbf{x}) - \nabla f^*(\mathcal{A}^\top \mathbf{y})\| \\
&\leq \|\mathcal{A}\| \cdot \tfrac{1}{\sigma} \cdot \|\mathcal{A}^\top \mathbf{x} - \mathcal{A}^\top \mathbf{y}\| \\
&\leq \|\mathcal{A}\| \cdot \tfrac{1}{\sigma} \cdot \|\mathcal{A}^\top\| \cdot \|\mathbf{x} - \mathbf{y}\| \\
&= \frac{\|\mathcal{A}\|^2}{\sigma} \|\mathbf{x} - \mathbf{y}\|
\end{aligned}$$

So $f^*(\mathcal{A}^\top \mathbf{y})$ is $\dfrac{\|\mathcal{A}\|^2}{\sigma}$-smooth.

# Solving the dual problem by proximal gradient method

- We just showed that $f^*(\mathcal{A}^\top \mathbf{y})$ is $\frac{\|\mathcal{A}\|^2}{\sigma}$-smooth.
- Hence for the dual problem

$$(D') \quad \min_{\mathbf{y}} q(\mathbf{y}) = f^*(\mathcal{A}^\top \mathbf{y}) + g^*(-\mathbf{y}),$$

  it is a minimization of a sum of smooth function and a possibly nonsmooth $g^*$, proximal gradient method can be used to solve $(D')$.
- That is, accelerated proximal gradient method (e.g. FISTA) can be used to solve $(D')$, with convergence rate of the sequence $\{\mathbf{y}_k\}$ towards the optimal point $\mathbf{y}^*$ (which exists due to duality, see last slide) as $\mathcal{O}(\frac{1}{k^2})$ :

$$q(\mathbf{y}^*) - q(\mathbf{y}_k) \leq \frac{2L\|\mathbf{y}_0 - \mathbf{y}^*\|^2}{k^2} \tag{1}$$

Small note : In FISTA, it is $F(\mathbf{x}_k) - F^* \leq \frac{2L\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k^2}$ , here it is $q(\mathbf{y}^*) - q(\mathbf{y}_k)$, the order is swapped since the problem on $\mathbf{y}$ is the dual problem, with the original form as $(D)$ which is maximization not minimization.

# The fast dual proximal gradient method

After some algebra to transform the first step of the fast proximal method in terms of the problem $(P)$, the iterations of the fast dual proximal gradient are

- $\mathbf{u}_k = \underset{\mathbf{x}}{\operatorname{argmax}} \ \langle \mathbf{x}, \mathbf{A}^\top \mathbf{w}_k \rangle - f(\mathbf{x})$
- $\mathbf{v}_k = \operatorname{prox}_{Lg}(\mathcal{A}\mathbf{u}_k - L\mathbf{w}_k)$
- $\mathbf{y}_k = \mathbf{w}_k - \frac{1}{L}(\mathcal{A}\mathbf{u}_k - \mathbf{v}_k)$
- $t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$
- $\mathbf{w}_{k+1} = \mathbf{y}_k + \frac{t_k - 1}{t_k}(\mathbf{y}_k - y_{k-1})$

where $L \geq \frac{\|\mathcal{A}\|^2}{\sigma}$, $\mathbf{w}_1 = \mathbf{y}_0$, $t_1 = 1$. Here the $\mathbf{y}_k$ is the sequence of dual varaible, $\mathbf{w}_k$ is the auxilary variable of the acceleration.

# Remarks

- The method is also an acceleration version of an aternating minimization scheme.
- The method is also an acceleration version of Uzawa method for minimizing a strongly convex function over a polyhedron.
- The convergence rate $\mathcal{O}(\frac{1}{k^2})$ is on the dual sequence $\{\mathbf{y}_k\}$, not the primal sequence $\{\mathbf{x}_k\}$
- As the algorithm only produces $\{\mathbf{y}_k\}$, to study the convergence of on $\{\mathbf{x}_k\}$, we define $\{\mathbf{x}_k\}$ from $\{\mathbf{y}_k\}$ as

$$\mathbf{x}_k = \operatorname*{argmax}_{\mathbf{x}} \langle \mathbf{x}, \mathcal{A}^\top \mathbf{y}_k \rangle - f(\mathbf{x}).$$

  Why define in this way : note that this is the formulation of convex conjugate we used to arrive the dual problem in slide 7. Previously we are given $\mathbf{x}$ and we want $\mathbf{y}$, now we are given $\mathbf{y}$ and we want $\mathbf{x}$.
- $\{\mathbf{x}_k\}$ is contained in dom$f$ but not necessarily feasible as $\mathcal{A}\mathbf{x}$ may not belong to dom$g$ (i.e. $\mathcal{A}\mathbf{x} \neq \mathbf{z}$). This infeasibility is common property of dual-based methods.

# Convergence rate on the primal sequence

**Theorem** Let $\{\mathbf{y}_k\}$ be the sequence produced by the fast dual proximal gradient method and let $\{\mathbf{x}_k\}$ be the primal sequence defined from $\{\mathbf{y}_k\}$ as

$$\mathbf{x}_k = \operatorname*{argmax}_{\mathbf{x}} \langle \mathbf{x}, \mathcal{A}^\top \mathbf{y}_k \rangle - f(\mathbf{x}).$$

Then

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq 2\sqrt{\frac{L}{\sigma}} \frac{\|\mathbf{y}_0 - \mathbf{y}^*\|}{k}.$$

Proof. For instrumental purpose, define

$$\begin{aligned} \mathbf{z}_k &\in \operatorname*{argmin}_{\mathbf{z} \in \mathrm{dom} g} \{\langle \mathbf{y}_k, \mathbf{z} \rangle + g(\mathbf{z})\} \\ h_1(\mathbf{x}) &= f(\mathbf{x}) - \langle \mathcal{A}^\top \mathbf{y}_k, \mathbf{x} \rangle \\ h_2(\mathbf{z}) &= g(\mathbf{z}) + \langle \mathbf{y}_k, \mathbf{z} \rangle \end{aligned}$$

Note that $L(\mathbf{x}, \mathbf{z}, \mathbf{y}_k) = h_1(\mathbf{x}) + h_2(\mathbf{z})$. Then by definitions of $\mathbf{x}_k$ and $\mathbf{z}_k$, we have

$$\begin{aligned} \mathbf{x}_k &= \operatorname*{argmin}_{\mathbf{x}} h_1(\mathbf{x}) \\ \mathbf{z}_k &= \operatorname*{argmin}_{\mathbf{z}} h_2(\mathbf{x}) \end{aligned}$$

# Convergence rate on the primal sequence

As $f$ is strongly convex, so as $h_1$. Hence on $h_1$ we have

$$h_1(\mathbf{x}) - h_1(\mathbf{x}_k) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|_2^2.$$

By definition of $\mathbf{z}_k$ we have for any $\mathbf{z} \in \text{dom}g$,

$$h_2(\mathbf{z}) - h_2(\mathbf{z}_k) \geq 0.$$

Sum the two inequalities gives

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2$$

Put $\mathbf{x} = \mathbf{x}^*$, $\mathbf{z}^* = \mathcal{A}\mathbf{x}^*$,

$$L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) \geq \frac{\sigma}{2}\|\mathbf{x} - \mathbf{x}_k\|^2$$

The goal is to get the bound

$$\|\mathbf{x} - \mathbf{x}_k\|^2 \leq \frac{2}{\sigma}\Big(L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k)\Big)$$

# Convergence rate on the primal sequence

Recall $q(\mathbf{y}) = -f^*(\mathcal{A}^\top \mathbf{y}) - g^*(-\mathbf{y})$

$$
\begin{aligned}
L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) &= \mathbf{h}_1(\mathbf{x}_k) + h_2(\mathbf{z}_k) \\
&= \underbrace{f(\mathbf{x}_k) - \langle \mathcal{A}^\top \mathbf{y}_k, \mathbf{x}_k \rangle}_{= -f^*(\mathcal{A}^\top \mathbf{y}_k)} \underbrace{+ g(\mathbf{z}_k) + \langle \mathbf{y}_k, \mathbf{z}_k \rangle}_{= -g^*(\mathbf{y}_k)} \\
&= q(\mathbf{y}_k) \\
L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*) &= \mathbf{h}_1(\mathbf{x}^*) + h_2(\mathbf{z}^*) \\
&= f(\mathbf{x}^*) - \langle \mathcal{A}^\top \mathbf{y}^*, \mathbf{x}^* \rangle + g(\mathbf{z}^*) + \langle \mathbf{y}^*, \mathbf{z}^* \rangle \\
&= f(\mathbf{x}^*) - \langle \mathbf{y}^*, \mathcal{A}\mathbf{x}^* \rangle + g(\mathbf{z}^*) + \langle \mathbf{y}^*, \mathbf{z}^* \rangle \\
&= f(\mathbf{x}^*) - \langle \mathbf{y}^*, \underbrace{\mathcal{A}\mathbf{x}^* - \mathbf{z}^*}_{= \mathbf{0}} \rangle + g(\mathbf{z}^*) \\
&= f(\mathbf{x}^*) + g(\mathcal{A}\mathbf{x}^*) \\
&= q(\mathbf{y}^*)
\end{aligned}
$$

where the last equality follows from strong duality. So we have

$$
\|\mathbf{x} - \mathbf{x}_k\|^2 \leq \frac{2}{\sigma} \Big( L(\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}_k) - L(\mathbf{x}_k, \mathbf{z}_k, \mathbf{y}_k) \Big) \leq \frac{2}{\sigma} \Big( q(\mathbf{y}^*) - q(\mathbf{y}_k) \Big)
$$

# Convergence rate on the primal sequence

Apply the FISTA convergence rate (1)

$$\|\mathbf{x}-\mathbf{x}_k\|^2 \leq \frac{2}{\sigma}\Big(q(\mathbf{y}^*)-q(\mathbf{y}_k)\Big) \overset{(1)}{\leq} \frac{2}{\sigma}\Big(\frac{2L\|\mathbf{y}_0-\mathbf{y}^*\|^2}{k^2}\Big) = \frac{4L}{\sigma}\Big(\frac{\|\mathbf{y}_0-\mathbf{y}^*\|}{k}\Big)^2$$

Ignore the middle

$$\|\mathbf{x}-\mathbf{x}_k\|^2 \leq \frac{4L}{\sigma}\Big(\frac{\|\mathbf{y}_0-\mathbf{y}^*\|}{k}\Big)^2$$

The square root on both side[2]

$$\|\mathbf{x}-\mathbf{x}_k\| \leq 2\sqrt{\frac{L}{\sigma}}\frac{\|\mathbf{y}_0-\mathbf{y}^*\|}{k}. \quad \square$$

Although we have a convergence rate, however, recall that $\{\mathbf{x}_k\}$ is contained in $\mathrm{dom}\, f$ but not necessarily feasible as $\mathcal{A}\mathbf{x}$ may not belong to $\mathrm{dom} g$ (i.e. $\mathcal{A}\mathbf{x} \neq \mathbf{z}$).

[2]Since $\|\mathbf{x}-\mathbf{x}_k\| \geq 0$ so taking square root does not change the inequlisty sign.

# Last page - summary

- Dual proximal gradient method
- Convergence rate of feasible dual sequence is $\mathcal{O}(\frac{1}{k^2})$

Reference

A. Beck and M. Teboulle, "A fast dual proximal gradient algorithm for convex minimization and applications", Operations Research Letters 2014

End of document