Primal-Dual Methods

V. Estellers

WS 2017

Recall: DUALITY

Theorem (Fenchel's Duality)

Let $G : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ and $F : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ be proper, closed, convex functions and $u \in$ ri$($dom$(G))$ such that $Ku \in$ ri$($dom$(F))$. Then

	$\inf_u$	$G(u) + F(Ku)$	*"Primal"*
$=$	$\inf_u \sup_q$	$G(u) + \langle q, Ku \rangle - F^*(q)$	*"Saddle point"*
$=$	$\sup_q \inf_u$	$G(u) + \langle q, Ku \rangle - F^*(q)$	*"Saddle point"*
$=$	$\sup_q$	$-G^*(-K^*q) - F^*(q)$	*"Dual"*

We used the dual formulation to solve problems of the form $\min_u \ \|u - f\|_2 + \alpha \|Du\|_1$ that we could not directly because the proximal operator of $\|Du\|_1$ is not simple.

Motivation

But we still do not have a method to solve problems of the form

$$\min_u \; \|u - f\|_1 + \alpha \|Du\|_1$$

although the proximal mapping of the ℓ^1-norm is easy to compute.

Can we build an algorithm around

$$\min_u \max_p G(u) + \langle p, Ku \rangle - F^*(p)?$$

Proximal mapping as implicit gradient descent

For differentiable E, the proximal mapping does an implicit gradient step

$$u^{k+1} = \mathsf{prox}_{\tau E}(u^k) \quad \Rightarrow u^{k+1} = u^k - \tau \nabla E(u^{k+1})$$

The primal-dual hybrid gradient algorithm

Let us define

$$\mathsf{PD}(u,p) := G(u) + \langle p, Ku\rangle - F^*(p)$$

and try to alternate implicit ascent steps in p with implicit descent steps in u:

$$\begin{aligned} p^{k+1} &= \ \mathsf{prox}_{-\sigma PD(u^k,\cdot)}(p^k) \\ u^{k+1} &= \ \mathsf{prox}_{\tau PD(\cdot,p^{k+1})}(u^k) \end{aligned}$$

One finds

$$\begin{aligned} p^{k+1} =& \mathsf{prox}_{-\sigma PD(u^k,\cdot)}(p^k), \\ =& \operatorname*{argmin}_p \frac{1}{2}\|p - p^k\|^2 + \sigma F^*(p) - \sigma\langle Ku^k, p\rangle \\ =& \operatorname*{argmin}_p \frac{1}{2}\|p - p^k - \sigma Ku^k\|^2 + \sigma F^*(p) \\ =& \mathsf{prox}_{\sigma F^*}(p^k + \sigma Ku^k) \end{aligned}$$

The primal-dual hybrid gradient algorithm

Let us define

$$\mathsf{PD}(u,p) := G(u) + \langle p, Ku\rangle - F^*(p)$$

and try to alternate implicit accent steps in p with implicit descent steps in u:

$$\begin{aligned} p^{k+1} &= \ \mathsf{prox}_{\sigma F^*}(p^k + \sigma K u^k) \\ u^{k+1} &= \ \mathsf{prox}_{\tau PD(\cdot, p^{k+1})}(u^k) \end{aligned}$$

One finds

$$\begin{aligned} u^{k+1} =& \mathsf{prox}_{\tau PD(\cdot, p^{k+1})}(u^k), \\ =& \operatorname*{argmin}_u \frac{1}{2}\|u - u^k\|^2 + \tau G(u) + \tau\langle Ku, p^{k+1}\rangle \\ =& \operatorname*{argmin}_u \frac{1}{2}\|u - u^k + \tau K^* p^{k+1}\|^2 + \tau G(u) \\ =& \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}) \end{aligned}$$

Primal-dual hybrid gradient method

We found

$$p^{k+1} = \mathrm{prox}_{\sigma F^*}(p^k + \sigma K u^k),$$
$$u^{k+1} = \mathrm{prox}_{\tau G}(u^k - \tau K^* p^{k+1}).$$

One should make one (currently non intuitive) modification:

Definition (PDHG)

We will call the iteration

$$\begin{aligned} p^{k+1} &= \mathrm{prox}_{\sigma F^*}(p^k + \sigma K \bar{u}^k), \\ u^{k+1} &= \mathrm{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned} \tag{PDHG}$$

the **Primal-Dual Hybrid Gradient Method**. As we will see, it converges if $\tau\sigma < \frac{1}{\|K\|^2}$.

References for PDHG

PDHG is commonly referred to as the Chambolle and Pock algorithm. Nevertheless, several authors contributed to its development.

Here is a (likely incomplete) list of relevant papers:

Pock, Cremers, Bischof, Chambolle, A convex relaxation approach for computing minimal partitions.

Esser, Zhang, Chan, A General Framework for a Class of First Order Primal-Dual Algorithms for Convex Optimization in Imaging Science.

Chambolle, Pock, A first-order primal-dual algorithm for convex problems with applications to imaging.

Zhang, Burger Osher, A unified primal-dual algorithm framework based on Bregman iteration.

Understanding PDHG

Why does PDHG work?

1. Sanity check: If the algorithm converges, it does so to a minimizer.
2. Why does PDHG converge? Computation on the board for

$$
\begin{aligned}
u^{k+1} &= \mathsf{prox}_{\tau G}(u^k - \tau K^* p^k) \\
\bar{u}^{k+1} &= 2u^{k+1} - u^k \\
p^{k+1} &= \mathsf{prox}_{\sigma F^*}(p^k + \sigma K \bar{u}^{k+1}).
\end{aligned} \tag{1}
$$

$$
\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}
$$

for the set-valued operator $T : \mathbb{R}^n \times \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^n) \times \mathcal{P}(\mathbb{R}^n)$

Fixed point iteration

We have reformulated the update rule to

$$0 \in Tz^{k+1} + M(z^{k+1} - z^k)$$

for a set-valued operator T and a matrix M. Let us define the process of computing the next iterate as the *resolvent*

$$z^{k+1} = (M + T)^{-1}(Mz^k). \tag{CPPA}$$

We already know an iteration of this form, the proximal point algorithm

$$u^{k+1} = prox_E(u^k) = (I + \tau\partial E)^{-1}(u^k)$$

So we can use the same tools to analyze its convergence. We will call it a *customized proximal point algorithm* (CPPA).

Convergence of the CPPA

Remember what we did for the proximal gradient algorithm?

$\rightarrow$ Show that $prox_E = (I + \tau \partial E)^{-1}$ is firmly nonexpansive, i.e. averaged with $\alpha = 1/2$.

We will do something similar by generalizing the crucial inequality

$$\langle p_u - p_v, u - v \rangle \geq 0 \qquad \forall u, v, p_u \in \partial E(u), p_v \in \partial E(v)$$

Definition (Monotone Operator)

A set valued operator T is called *monotone* if the inequality

$$\langle p_u - p_v, u - v \rangle \geq 0$$

holds for all u, v, $p_u \in T(u)$ and $p_v \in T(v)$.

Convergence of the CPPA

This has the potential to show convergence of

$$0 \in T(z^{k+1}) + z^{k+1} - z^k, \tag{PPA}$$

provided that the above iteration is well-defined, i.e. the resolvent $(I + T)^{-1}(z)$ is defined for any $z \in \mathbb{R}^n$. This is a technical issue which can be resolved by considering *maximal* monotone operators. In our convex settings, this is not an issue.

Definition

The relation T is **maximal monotone** if there is no monotone operator that properly contains it as a subset of $\mathbb{R}^n \times \mathbb{R}^n$.

In other words, if the monotone operator T is not maximal, then there is $(x, u) \notin T$ such that $T \cup \{(x, u)\}$ is monotone.

Examples of maximal monotone operators

Lemma

$E\colon \mathbb{R}^n \to \bar{\mathbb{R}}$, then ∂E is a monotone operator. If E is closed convex and proper then ∂E is maximal monotone.

Lemma

A continuous monotone function $F\colon \mathbb{R}^n \to \mathbb{R}$ with dom$(F) = \mathbb{R}^n$ is maximal.

Lemma

If T is maximal monotone, then the resolvent $R_T = (I + \alpha T)^{-1}$ with $\alpha > 0$ and the Caley operator $C_T = 2R_T - I$ are nonexpansive functions.

Theorem (Convergence of Generalized Proximal Point Algorithm)

Let T be a maximal monotone operator, and let there exist a z such that $0 \in T(z)$. Then the (generalized) proximal point algorithm

$$\begin{aligned} z^{k+1} &= (T+I)^{-1}(z^k) \\ 0 &\in T(z^{k+1}) + z^{k+1} - z^k \end{aligned} \tag{2}$$

converges to a point $\tilde{z}$ with $0 \in T(\tilde{z})$.

Proof.

If T is maximal monotone, the resolvent $R_T = (T+I)^{-1}$ and the Caley operator $C_T = 2R_T - I$ are nonexpansize. Since $R_T = \frac{1}{2}I + \frac{1}{2}C_T$, the resolvent R_T is an averaged operator and the generalized proximal point algorithm is a fixed-point iteration of an averaged operator that converges by Krasnoselskii-Mann Theorem. $\square$

Convergence of the CPPA

But we wrote the PDHG algorithm as

$$0 \in T(z^{k+1}) + Mz^{k+1} - Mz^k, \tag{3}$$

i.e. with an additional matrix M.
Idea: For symmetric positive definite matrices, write $M = L^T L$ and rewrite (CPPA) as

$$0 \in L^{-T} T L^{-1}(\zeta^{k+1}) + \zeta^{k+1} - \zeta^k, \tag{CPPA}$$

with $\zeta^k = Lz^k$, and

$$L^{-T} T L^{-1}(\zeta) = \{q \in \mathbb{R}^n \mid q = L^{-T} p, \;\; p \in T(L^{-1}\zeta)\}.$$

Lemma

If T is monotone, then $L^{-T} T L^{-1}$ is monotone, too.

Proof: Exercise.

Convergence conclusions CPPA

Theorem (Convergence CPPA)

Let T be a maximally monotone operator. Let there exist a z such that $0 \in T(z)$, and let the matrix M be symmetric positive definite. Then the customized proximal point algorithm

$$z^{k+1} = (M + T)^{-1}(Mz^k)$$

converges to a $\hat{z}$ with $0 \in T(z)$.

Convergence conclusions PDHG

As the primal-dual hybrid gradient method can be rewritten (after an index shift) as

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \underbrace{\begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix}}_{=:T} \begin{pmatrix} u^{k+1} \\ p^{k+1} \end{pmatrix} + \underbrace{\begin{pmatrix} \frac{1}{\tau} I & -K^T \\ -K & \frac{1}{\sigma} I \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{k+1} - p^k \end{pmatrix}.$$

Theorem

Convergence PDHG The operator T is maximally monotone. For $\tau\sigma < \frac{1}{\|K\|^2}$ the matrix M in the PDHG algorithm is positive definite. Hence, PDHG converges.

(Assuming F and G to be proper, closed, and convex, assuming there is a $u \in$ ri(G) such that $Ku \in$ ri(F), and assuming the existence of a minimizer).

ROF Denoising

$$\min P(u) = \min_u \frac{1}{2}\|u - f\|^2 + \alpha\|Ku\|_{2,1}$$

with K being a discretization of the multichannel gradient operator.

ROF Denoising

We write

$$\min_u P(u) = \min_u \max_p \frac{1}{2}\|u - f\|^2 + \langle Ku, p\rangle - \iota_{\|\cdot\|_{2,\infty} \le \alpha}(p).$$

The (PDHG) updates are

$$\begin{aligned} p^{k+1} &= \mathsf{prox}_{\sigma F^*}(p^k + \sigma K \bar{u}^k) \\ u^{k+1} &= \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned}$$

which in this case amounts to

$$\begin{aligned} p^{k+1} &= \operatorname*{argmin}_p \frac{1}{2}\|p - (p^k + \sigma K \bar{u}^k)\|^2 + \sigma \iota_{\|\cdot\|_{2,\infty} \le \alpha}(p), \\ u^{k+1} &= \operatorname*{argmin}_u \frac{1}{2}\|u - (u^k - \tau K^* p^{k+1})\|^2 + \frac{\tau}{2}\|u - f\|^2 \\ &= \frac{u^k - \tau K^* p^{k+1} + \tau f}{1 + \tau} \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned}$$

TV-L^1 Denoising

$$\min P(u) = \min_u \|u - f\|_1 + \alpha \|Ku\|_{2,1}$$

with K being a discretization of the multichannel gradient operator.

TV-L^1 Denoising

We write

$$\min_u P(u) = \min_u \max_p \frac{1}{2}\|u-f\|_1 + \langle Ku, p\rangle - \iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p).$$

The (PDHG) updates are

$$\begin{aligned} p^{k+1} &= \mathsf{prox}_{\sigma F^*}(p^k + \sigma K\bar{u}^k) \\ u^{k+1} &= \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned}$$

which in this case amounts to

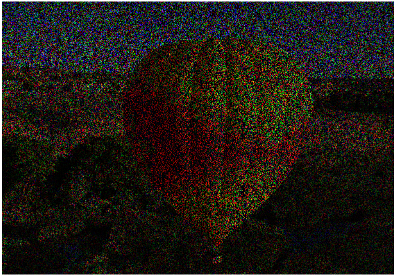
An exercise! :-)

TV-Inpainting

$$\min P(u) = \min_u \iota_{f_{|I}}(u) + \alpha \|Ku\|_{2,1}$$

with K being a discretization of the color gradient operator, and

$$\iota_{f_{|I}}(u) = \begin{cases} 0 & \text{if } u_i = f_i \text{ for all } i \in I, \\ \infty & \text{otherwise.} \end{cases}.$$



TV-Inpainting

We write

$$\min_u P(u) = \min_u \max_p \iota_{f_{|I}}(u) + \langle Ku, p \rangle - \iota_{\|\cdot\|_{2,\infty} \leq \alpha}(p).$$

The (PDHG) updates are

$$\begin{aligned}
p^{k+1} &= \mathrm{prox}_{\sigma F^*}(p^k + \sigma K \bar{u}^k) \\
u^{k+1} &= \mathrm{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \\
\Rightarrow u_i^{k+1} &= \begin{cases} f_i & \text{if } i \in I, \\ (u^k - \tau K^* p^{k+1})_i & \text{otherwise.} \end{cases} \\
\bar{u}^{k+1} &= 2u^{k+1} - u^k.
\end{aligned}$$

TV-Deblurring

$$\min P(u) = \min_u \frac{1}{2}\|Au - f\|^2 + \alpha\|Ku\|_{2,1}$$

with K being a discretization of the multichannel gradient operator, A being a convolution with a blur kernel.

TV-Deblurring - Option 1

We write

$$\min_u P(u) = \min_u \max_p \frac{1}{2}\|Au - f\|^2 + \langle Ku, p\rangle - \iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p).$$

The (PDHG) updates are

$$\begin{aligned} p^{k+1} &= \mathsf{prox}_{\sigma F^*}(p^k + \sigma K\bar{u}^k) \\ u^{k+1} &= \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned}$$

which in this case amounts to

$$\begin{aligned} p^{k+1} &= \operatorname*{argmin}_p \frac{1}{2}\|p - (p^k + \sigma K\bar{u}^k)\|^2 + \sigma\iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p), \\ u^{k+1} &= \operatorname*{argmin}_u \frac{1}{2}\|u - (u^k - \tau K^* p^{k+1})\|^2 + \frac{\tau}{2}\|Au - f\|^2 \\ &= (I + \tau A^*A)^{-1}(u^k - \tau K^* p^{k+1} + \tau f) \\ \bar{u}^{k+1} &= 2u^{k+1} - u^k. \end{aligned}$$

TV-Deblurring - Option 2

We write

$$
\begin{aligned}
&\min_u P(u) \\
&= \min_u \max_{p,q} \langle Au - f, q\rangle - \frac{1}{2}\|q\|^2 + \langle Ku, p\rangle - \iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p) \\
&= \min_u \max_{p,q} \left\langle \begin{pmatrix} A \\ K \end{pmatrix} u, \begin{pmatrix} q \\ p \end{pmatrix} \right\rangle - \langle f, q\rangle - \frac{1}{2}\|q\|^2 - \iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p)
\end{aligned}
$$

Now we have

$$
\begin{aligned}
F^*(p,q) &= \langle f, q\rangle + \frac{1}{2}\|q\|^2 + \iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p) \\
G(u) &= 0 \\
\tilde{K} &= \begin{pmatrix} A \\ K \end{pmatrix}
\end{aligned}
$$

TV-Deblurring - Option 2

The (PDHG) updates are

$$
\begin{aligned}
q^{k+1} &= \operatorname*{argmin}_q \frac{1}{2}\|q - (q^k + \sigma A\bar{u}^k)\|^2 + \sigma\langle f, q\rangle + \frac{\sigma}{2}\|q\|^2, \\
p^{k+1} &= \operatorname*{argmin}_p \frac{1}{2}\|p - (p^k + \sigma K\bar{u}^k)\|^2 + \sigma\iota_{\|\cdot\|_{2,\infty}\leq\alpha}(p), \\
u^{k+1} &= u^k - \tau K^* p^{k+1} - \tau A^* q^{k+1} \\
\bar{u}^{k+1} &= 2u^{k+1} - u^k.
\end{aligned}
$$

TV-Zooming

$$\min P(u) = \min_u \frac{1}{2}\|Au - f\|^2 + \alpha \|Ku\|_{2,1}$$

with K being a discretization of the multichannel gradient operator, $A = DB$, with B being a convolution with a blur kernel, and D being a downsampling, e.g. a matrix

$$D = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & \ldots & \ldots \\ 0 & 0 & 1 & 0 & 0 & \ldots & \ldots \\ 0 & 0 & 0 & 0 & 1 & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \end{pmatrix}$$

PDHG implementation: Option 2 from the previous example.

TV-Zooming

Input data

Nearest neighbor

TV Zooming

Image Segmentation

$$\min P(u) = \min_u \iota_\Delta(u) + \iota_{\geq 0}(u) + \langle u, f \rangle + \alpha \|Ku\|_{2,1}$$

where $K : \mathbb{R}^{n\times m\times c} \to \mathbb{R}^{nmc\times 2}$ being a discretization of the multichannel gradient operator, and

$$\iota_\Delta(u) = \begin{cases} 0 & \text{if } \sum_k u_{i,j,k} = 1, \ \forall (i,j) \\ \infty & \text{else.} \end{cases}$$

$$\iota_{\geq 0}(u) = \begin{cases} 0 & \text{if } u_{i,j,k} \geq 0, \ \forall (i,j,k) \\ \infty & \text{else.} \end{cases}$$

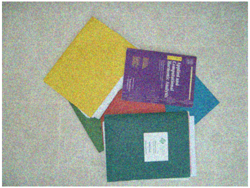
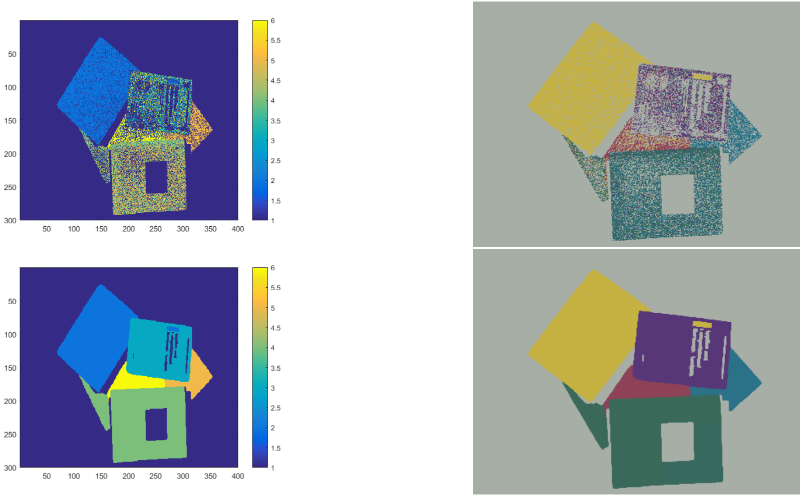


Image Segmentation



Upper row: data term minimization (=kmeans assignment), lower row: variational method

Image Segmentation

Option 1: We solve

$$\min_{u} \max_{p} \iota_{\Delta}(u) + \iota_{\geq 0}(u) + \langle u, f \rangle + \langle Ku, p \rangle - \iota_{\|\cdot\|_{2,\infty} \leq \alpha}(p).$$

$\rightarrow$ Primal proximal operator: Projection onto unit simplex.

Option 2: We solve

$$\min_{u} \max_{p.q} \langle Su - 1, q \rangle + \iota_{\geq 0}(u) + \langle u, f \rangle + \langle Ku, p \rangle - \iota_{\|\cdot\|_{2,\infty} \leq \alpha}(p).$$

where $(Su)_{i,j} = \sum_k u_{i,j}$.

$\rightarrow$ Very simple proximal operators, but additional variable.

Final remark for applications

If you are too lazy to compute the proximity operator of F^*

$$
\begin{aligned}
\tilde{p} &= \mathrm{prox}_{\sigma F^*}(z) \\
&= \arg\min_p \frac{1}{2}\|p - z\|^2 + \sigma F^*(p) \\
\Rightarrow 0 &= \tilde{p} - z + \sigma \tilde{u}, \quad \tilde{u} \in \partial F^*(\tilde{p}) \\
\Rightarrow 0 &= \tilde{u} - z/\sigma + \frac{1}{\sigma}\tilde{p}, \quad \tilde{p} \in \partial F(\tilde{u}) \\
\Rightarrow \tilde{u} &= \mathrm{prox}_{\frac{1}{\sigma}F}(z/\sigma) \\
\Rightarrow \tilde{p} &= z - \sigma \ \mathrm{prox}_{\frac{1}{\sigma}F}(z/\sigma)
\end{aligned}
$$

Lemma (Moreau's identity)

If you know prox_F *you also know* prox_{F^*},

$$
\mathrm{prox}_{\sigma F^*}(z) = z - \sigma \ \mathrm{prox}_{\frac{1}{\sigma}F}(z/\sigma).
$$

Convergence rate

We have seen: PDHG works very well on problems of the form

$$\min G(u) + F(Ku),$$

for which F and G are simple.

We get a convergence rate of

$$\min_{j \in \{0,\dots,k\}} \|(I + L^{-T} T L^{-1})(\xi^k) - \xi^k\|^2 \leq C \frac{\|\xi^0 - \xi^0\|}{k+1}$$

for $\xi^k = L(u^k, p^k)$, L being the matrix square-root of M, and C being a constant.

What if our problem is more friendly? E.g. what if G or F or both are strongly convex?

Either G or F^* is strongly convex

$$
\begin{aligned}
p^{k+1} &= \mathrm{prox}_{\sigma_k F^*}(p^k + \sigma_k K \bar{u}^k), \\
u^{k+1} &= \mathrm{prox}_{\tau_k G}(u^k - \tau_k K^* p^{k+1}), \\
\theta_k &= \frac{1}{\sqrt{1 + 2\gamma\tau_k}}, \\
\tau_{k+1} &= \theta_k \tau_k, \quad \sigma_{k+1} = \sigma_k / \theta_k \\
\bar{u}^{k+1} &= u^{k+1} + \theta_k (u^{k+1} - u^k).
\end{aligned}
\tag{PDHG2}
$$

for $\tau_0 \sigma_0 \leq \|K\|^2$, and G being γ-strongly convex.

Theorem

For strongly convex G and $\epsilon > 0$, there exists an N_0 such that for any $N \geq N_0$:

$$
\|\tilde{u} - u^N\|^2 \leq \frac{1+\epsilon}{\gamma^2 N^2} \left(\frac{\|\tilde{u} - u^0\|^2}{\tau_0^2} + \|K\|^2 \|\tilde{p} - p^0\|^2 \right)
$$

Discussion of the convergence orders

If part of the energy is L smooth, the gradient methods obtain linear convergence on strongly convex energies.

As L-smoothness of the primal corresponds to $1/L$-strong convexity of the convex conjugate. It is natural to ask: what can we do if we additionally assume F to be L-smooth, i.e., assume F^* to be strongly convex?

Two strongly convex functions

Consider

$$p^{k+1} = \mathsf{prox}_{\sigma F^*}(p^k + \sigma K \bar{u}^k),$$
$$u^{k+1} = \mathsf{prox}_{\tau G}(u^k - \tau K^* p^{k+1}), \qquad \text{(PDHG3)}$$
$$\bar{u}^{k+1} = u^{k+1} + \theta(u^{k+1} - u^k).$$

Theorem (Linear convergence of strongly convex functions)

For $\mu \leq 2\sqrt{\gamma\delta}/\|K\|$, $\tau = \mu/(2\gamma)$, $\sigma = \mu/(2\delta)$, $\theta \in [1/(1+\mu), 1]$, G *being* γ*-strongly convex and* F^* *being* δ*-strongly convex, there exists* $\omega < 1$, *such that the iterates of* (PDHG3) *meet*

$$\gamma\|u^N - \tilde{u}\|^2 + (1-\omega)\delta\|p^N - \tilde{p}\|^2 \leq \omega^N(\gamma\|u^0 - \tilde{u}\|^2 + \delta\|p^0 - \tilde{p}\|^2).$$

Generic form

Remember the optimality conditions of the saddle point formulation

$$\min_u \max_p \; G(u) + \langle Ku, p\rangle - F^*(p)$$

were

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{p} \end{pmatrix}.$$

We could not compute $(\hat{u}, \hat{p})$ directly. Therefore,

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix} \begin{pmatrix} u^{k+1} \\ p^{K+1} \end{pmatrix} + \underbrace{\begin{pmatrix} M_1 & M_3 \\ M_4 & M_2 \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{K+1} - p^k \end{pmatrix}$$

such that

- M is symmetric, i.e. $M_3 = (M_4)^T$,
- sequential updates are possible, i.e. $M_3 = -K^T$, or $M_4 = K$.

Diagonal M_1 and M_2

Sticking to $M_3 = -K^T$ leads to

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial G & K^T \\ -K & \partial F^* \end{pmatrix} \begin{pmatrix} u^{k+1} \\ p^{K+1} \end{pmatrix} + \underbrace{\begin{pmatrix} M_1 & -K^T \\ -K & M_2 \end{pmatrix}}_{=:M} \begin{pmatrix} u^{k+1} - u^k \\ p^{K+1} - p^k \end{pmatrix}.$$

Only remaining requirement: M should be positive definite!

In PDHG we chose $M_1 = \frac{1}{\tau} I$, $M_2 = \frac{1}{\sigma} I$ for simplicity.

In many cases, e.g., for separable F^* and G, the updates remain easy to compute if M_1 and M_2 are diagonal.

Theorem

Let $\alpha \in [0, 2]$, $M_1 =$ *diag*(m_j^1) *and* $M_2 =$ *diag*(m_i^2) *with*

$$m_j^1 > \sum_i |K_{i,j}|^{2-\alpha}, \qquad m_i^2 > \sum_j |K_{i,j}|^{\alpha}.$$

Then M *is positive definite.*

Some remarks

Regarding the choice of M_1 and M_2:

- It does not influence the convergence rate.
- It is an active field of research to understand its influence on constants in the convergence rates.
- It can make a huge difference in practice!!
- Typically, the practical convergence speed improves the more information about K is included in M_1, M_2.

The latter motivates yet a different and vastly popular algorithm, the **alternating direction method of multipliers (ADMM)**.