Continuous (convex) optimisation

M2 - PSL / Dauphine / S.U.

Antonin Chambolle, CNRS, CEREMADE

Université Paris Dauphine PSL

Oct.-Dec. 2021

Lecture 5: Saddle points, Primal-dual splitting

Contents

Constrained problems. Duality.

Assume we need to solve:

$$min_x \{f(x) : g_i(x) \leq 0, i = 1, \dots, m\}$$

with f, g_i convex (KKT framework), and we assume in addition:

- f is strongly convex with some parameter $\gamma > 0$,
- $|g(x) - g(x')| \leq L|x - x'|$ ($g = (g_1, \dots, g_m)$) is L-Lipschitz).

We can introduce a Lagrange multiplier for the constraints as in the KKT's theorem:

$$\min_{Ax \leq b} f(x) = \min_x \sup_{\lambda \geq 0} f(x) + \langle \lambda, g(x) \rangle = (\geq) \sup_{\lambda \geq 0} \min_x f(x) + \langle \lambda, g(x) \rangle$$

and try to solve the *dual problem*

$$\max_{\lambda \geq 0} \mathcal{D}(\lambda) \quad \text{where} \quad \mathcal{D}(\lambda) = \min_x f(x) + \langle \lambda, g(x) \rangle .$$

Uzawa

Assume now we are able to solve for any $\lambda \geq 0$ the unconstrained problem

$$\min_x f(x) + \langle \lambda, g(x) \rangle$$

(for instance, using FISTA...)
Let $x(\lambda)$ be the (unique) solution. Then for any $\mu \geq 0$,

$$\begin{aligned}
\mathcal{D}(\mu) = f(x(\mu)) + \langle \mu, g(x(\mu)) \rangle &= f(x(\mu)) + \langle \lambda, g(x(\mu)) \rangle + \langle \mu - \lambda, g(x(\mu)) \rangle \\
&\geq f(x(\lambda)) + \langle \lambda, g(x(\lambda)) \rangle + \frac{\gamma}{2}|x(\mu) - x(\lambda)|^2 + \langle \mu - \lambda, g(x(\mu)) \rangle ,
\end{aligned}$$

that is:

$$\mathcal{D}(\lambda) \leq \mathcal{D}(\mu) + \langle \lambda - \mu, g(x(\mu)) \rangle - \frac{\gamma}{2}|x(\mu) - x(\lambda)|^2$$

and it follows:

Uzawa

$$\mathcal{D}(\lambda) \le \mathcal{D}(\mu) + \langle \lambda - \mu, g(x(\mu)) \rangle - \frac{\gamma}{2}|x(\mu) - x(\lambda)|^2$$

and it follows:

$$g(x(\mu)) \in \partial\mathcal{D}(\mu)$$

[here the *supergradient* of the concave function g] and

$$\gamma|x(\mu) - x(\lambda)|^2 \le \langle \lambda - \mu, g(x(\mu)) - g(x(\lambda)) \rangle \le |\lambda - \mu||g(x(\mu)) - g(x(\lambda))|.$$

Now we have $|g(x(\mu)) - g(x(\lambda))| \le L|x(\mu) - x(\lambda)|$ and we deduce

$$|g(x(\mu)) - g(x(\lambda))| \le \frac{L^2}{\gamma}|\lambda - \mu|$$

that is, $\mathcal{D}$ is concave with L^2/γ-Lipschitz gradient.

Uzawa

Then, it can be solved using “ISTA” or “FISTA”, for instance:

$$\lambda^{k+1} = \left(\lambda^k + \tau g(x(\lambda^k))\right)^+$$

for $\tau = \gamma/L^2$, which will ensure that:

$$\mathcal{D}(\lambda^*) - \mathcal{D}(\lambda^N) \leq \frac{L^2}{2\gamma N}|\lambda^0 - \lambda^*|^2.$$

In addition (using $\mu = \lambda^*$ in the first inequality of the previous slide),

$$|x(\lambda^N) - x^*|^2 \leq \frac{2}{\gamma}\left(\mathcal{D}(\lambda^*) - \mathcal{D}(\lambda^N)\right) \leq \frac{L^2}{\gamma^2 N}|\lambda^0 - \lambda^*|^2.$$

(Of course, one should use acceleration, but for this we need to be able to solve the primal problems very precisely.)

The "ADMM"

or Alternating Directions Method of Multipliers

The "ADMM" aims at solving a slightly more general form than $f(Kx)+g(x)$, namely:

$$\min_{Ax+By=b} f(x)+g(y) \tag{1}$$

for f,g convex, lsc., and A,B continuous, linear operators. [Of course, it is still of the form $f(Kx)+g(x)$ for some other functions f,g, which?]

It has the dual form:

$$\max_p \langle b,p\rangle - f^*(A^*p) - g^*(B^*p)$$

with strong duality if f,g are continuous at some x,y with $Ax+By=b$ (in finite dimension, x,y in the relative interiors of the domains, respectively, of f,g) or if f^* is continuous at some point A^*p and g^* at B^*p (in finite dimension, $A^*p\in \operatorname{ri}\operatorname{dom} f^*$, $B^*p\in \operatorname{ri}\operatorname{dom} g^*$ for some p). This seems not particularly easier to solve for generic f,g.

ADMM: Augmented Lagrangian

An "augmented Lagrangian" approach consists in introducing the constraint in the form

$$\min_{x,y} \sup_{z} f(x) + g(y) - \langle z, Ax + By - b\rangle + \frac{\gamma}{2}|Ax + By - b|^2$$

for some $\gamma > 0$, which is equivalent (as the sup is $+\infty$ if $Ax + By \neq b$) to the original problem. Why use $\gamma > 0$? It makes the problem more regular.

ADMM: Augmented Lagrangian and dual

One considers the dual (concave) function:

$$\mathcal{D}(z) = \inf_{x,y} f(x) + g(y) - \langle z, Ax + By - b\rangle + \frac{\gamma}{2}|Ax + By - b|^2$$

Thanks to the quadratic term, it has $(1/\gamma)$-Lipschitz gradient. This follows from the following result which we will prove next week in a slightly more general setting:

Lemma

Let f be convex, lsc: then f is γ-convex (strongly convex with parameter γ) if and only if f^* has $(1/\gamma)$-Lipschitz gradient.

ADMM: Augmented Lagrangian and dual

Hence, a natural method for maximizing the dual could be to implement an (accelerated) gradient ascent, using that (the supergradient)

$$\partial\mathcal{D}(z) = \{-(Ax + By - b)\}$$

where (x, y) minimizes the problem which defines $\mathcal{D}(z)$. (Same proof as for the Uzawa method, or simply Legendre-Fenchel identity.)

However, it means we are able to solve for (x, y), which is not necessarily easy. Hence the “Alternating Directions Methods of Multipliers”.

ADMM: algorithm

[Proposed initially by Glowinski and Marroco 75 / Gabay and Mercier 76]

Choose $\gamma > 0$, y^0, z^0.
for all $k \geq 0$ **do**
Find x^{k+1} by minimising $x \mapsto f(x) - \left\langle z^k, Ax \right\rangle + \frac{\gamma}{2}|b - Ax - By^k|^2$,
Find y^{k+1} by minimising $y \mapsto g(y) - \left\langle z^k, By \right\rangle + \frac{\gamma}{2}|b - Ax^{k+1} - By|^2$,
Update $z^{k+1} = z^k + \gamma(b - Ax^{k+1} - By^{k+1})$.
end for

Convergence: for f, g convex, lsc. and provided there exists a saddle-point, the method converges.

Proof is omitted. In fact, it can be related to a Douglas-Rachford iteration on the dual problem. Or it is an "inexact" gradient ascent on the dual, with an error which needs to be controlled.

ADMM: difficulties

In practice, it is not necessarily easy to solve

$$\min_x f(x) - \left\langle z^k, Ax \right\rangle + \frac{\gamma}{2}|b - Ax - By^k|^2$$

and one may revert to "proximal" ADMM: one introduces G, H symmetric positive-definite operators and considers rather the steps:

$$x^{k+1} = \arg\min_x f(x) - \left\langle z^k, Ax \right\rangle + \frac{\gamma}{2}|b - Ax - By^k|^2 + \frac{1}{2}|x - x^k|_F^2,$$
$$y^{k+1} = \arg\min_y g(y) - \left\langle z^k, By \right\rangle + \frac{\gamma}{2}|b - Ax^{k+1} - By|^2 + \frac{1}{2}|y - y^k|_G^2.$$

In practice, choosing $F = I/\tau - \gamma A^*A$ and $G = I/\sigma - \gamma B^*B$ with τ, σ small enough allows to solve the problems if the "prox" of f, g can be computed. Then, again, the algorithm will converge.

"PDHG"
(primal dual hybrid gradient)

One considers again:

$$\min_x f(Kx) + g(x) = \min_x \sup_y \langle Kx, y\rangle + g(x) - f^*(y).$$

A basic idea consists in performing a gradient descent in x and a gradient ascent in y ("Arrow-Hurwicz" method):

$$x^{k+1} = (I + \tau\partial g)^{-1}(x^k - \tau K^* y^k)\,,$$
$$y^{k+1} = (I + \sigma\partial f^*)^{-1}(y^k + \sigma K x^k)$$

for some $\sigma, \tau > 0$, however in general this will not converge (case $f, g = 0$: this is similar to an explicit update for a monotone operator).

We observe though that in this specific case, one could use x^{k+1} in the second step ($\to$ semi implicit). Does it help?

PDHG

Well, almost. For $f, g = 0$ one has:

$$\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} = \begin{pmatrix} I & -\tau K^* \\ \sigma K & I - \sigma\tau KK^* \end{pmatrix} \begin{pmatrix} x^k \\ y^k \end{pmatrix}$$

and the eigenvalues of this matrix have modulus equal to 1 for $\sigma\tau$ small enough.

PDHG

Well, almost. For $f, g = 0$ one has:

$$\begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} = \begin{pmatrix} I & -\tau K^* \\ \sigma K & I - \sigma\tau KK^* \end{pmatrix} \begin{pmatrix} x^k \\ y^k \end{pmatrix}$$

and the eigenvalues of this matrix have modulus equal to 1 for $\sigma\tau$ small enough.

We write, for $\lambda \in \mathbb{C}$,

$$\lambda \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} I & -\tau K^* \\ \sigma K & I - \sigma\tau KK^* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \Leftrightarrow \begin{cases} x = \frac{\tau}{1-\lambda} K^* y & \text{or } \lambda = 1,\ K^* y = 0 \\ \sigma K x + y - \sigma\tau KK^* y = \lambda y \end{cases}$$

In case $\lambda = 1$ we also deduce that $Kx = 0$. So the eigenvalue 1 corresponds to $x \in \ker K$, $y \in \ker K^*$. If $K \neq 0$ there must be another eigenvalue $\lambda \neq 1$. Then, one has:

$$\frac{\sigma\tau}{1-\lambda} KK^* y - \sigma\tau KK^* y = (\lambda - 1) y \Leftrightarrow KK^* y = -\frac{(\lambda-1)^2}{\sigma\tau\lambda} y.$$

unless $\lambda = 0$ but in this case $y = 0$, then $x = 0$, and it is not an eigenvalue.

PDHG

We see that y is an eigenvector of KK^*, corresponding to an eigenvalue $\mu > 0$ (otherwise $\lambda = 1$). λ solves:

$$-\frac{(\lambda-1)^2}{\sigma\tau\lambda} = \mu \Leftrightarrow \lambda^2 - 2\lambda + 1 = -\sigma\tau\mu\lambda \Leftrightarrow \lambda^2 - 2(1 - \tfrac{\sigma\tau\mu}{2})\lambda + 1 = 0$$

If $\sigma\tau\|K^*K\| \leq 2$, letting $1 - \sigma\tau\mu/2 = \cos\theta$ we find that $\lambda = \cos\theta \pm i\sin\theta$.

Hence, in that case, the algorithm will not converge nor diverge (the iterates "rotate"). Of course, for $f, g \neq 0$, the method may actually converge, in practice.

PDHG

The PDHG algorithm is a stable and converging variant of the previous case. Its simplest form is:

$$\begin{aligned} x^{k+1} &= (I+\tau\partial g)^{-1}(x^k - \tau K^* y^k)\,, \\ y^{k+1} &= (I+\sigma\partial f^*)^{-1}(y^k + \sigma K(2x^{k+1} - x^k)) \end{aligned} \qquad (PDHG)$$

Proposition (He-Yuan 2011)

If $\tau\sigma\|K^*K\| < 1$ then PDHG[a] is a proximal-point algorithm.

[a] "Primal-dual hybrid gradient"

PDHG

To see this we write the iterates as follows:

$$\begin{cases} \frac{x^{k+1}-x^k}{\tau} + \partial g(x^{k+1}) \ni -K^*y^k = K^*(y^{k+1} - y^k) - K^*y^{k+1} \\ \frac{y^{k+1}-y^k}{\sigma} + \partial f^*(y^{k+1}) \ni K(x^{k+1} - x^k) + Kx^{k+1}, \end{cases}$$

that is

$$\begin{pmatrix} \frac{1}{\tau}I & -K^* \\ -K & \frac{1}{\sigma}I \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix} + \begin{pmatrix} \partial g(x^{k+1}) \\ \partial f^*(y^{k+1}) \end{pmatrix} + \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} \ni 0.$$

PDHG

We remark that if S is symmetric, positive-definite (defines a metric/coercive in infinite dimension) then for A a maximal monotone operator:

$$S(z^{k+1} - z^k) + Az^{k+1} \ni 0$$

is the iteration of the proximal point algorithm for the maximal monotone operator $S^{-1}A$ in the metric defined by the scalar product $\langle z, z' \rangle_S := \langle Sz, z' \rangle$.

PDHG

Hence here, one find that the algorithm is a PPA iff

$$M_{\tau,\sigma} := \begin{pmatrix} \frac{1}{\tau} I & -K^* \\ -K & \frac{1}{\sigma} I \end{pmatrix}$$

is symmetric, coercive.

PDHG

One has:

$$\left\langle M_{\tau,\sigma}\begin{pmatrix}\xi\\ \eta\end{pmatrix},\begin{pmatrix}\xi\\ \eta\end{pmatrix}\right\rangle = \tfrac{1}{\tau}|\xi|^2 + \tfrac{1}{\sigma}|\eta|^2 - 2\langle K\xi,\eta\rangle$$

is positive if and only if for any $X, Y \geq 0$

$$\sup_{|\xi|\leq X,|\eta|\leq Y} 2\langle K\xi,\eta\rangle = 2\|K\|XY < \frac{X^2}{\tau} + \frac{Y^2}{\sigma}$$

if and only if

$$2\|K\| < \min_{X\geq 0, Y\geq 0} \frac{X}{\tau Y} + \frac{Y}{\sigma X} = \frac{2}{\sqrt{\tau\sigma}}$$

if and only if

$$\tau\sigma\|K\|^2 < 1,$$

hence the theorem.

PDHG: rate

One inherits the rate of convergence for the iterates of a proximal-point algorithm.
Yet for this specific form (using the convexity of f^*, g) one can improve the rate.
We denote $z = (x, y)^T$ and take the scalar product of the algorithm and $z^{k+1} - z$:

$$\left\langle z^{k+1} - z^k, z^{k+1} - z \right\rangle_{M_{\tau,\sigma}} + \left\langle \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix}, \begin{pmatrix} x^{k+1} - x \\ y^{k+1} - y \end{pmatrix} \right\rangle \\ + g(x^{k+1}) + f^*(y^{k+1}) \le g(x) + f^*(y)$$

The scalar product is

$$-\left\langle K^* y^{k+1}, x \right\rangle + \left\langle K x^{k+1}, y \right\rangle$$

while

$$\left\langle z^{k+1} - z^k, z^{k+1} - z \right\rangle_{M_{\tau,\sigma}} = \frac{1}{2}|z^{k+1} - z^k|^2_{M_{\tau,\sigma}} + \frac{1}{2}|z^{k+1} - z|^2_{M_{\tau,\sigma}} - \frac{1}{2}|z^k - z|^2_{M_{\tau,\sigma}}.$$

Hence:

$$\frac{1}{2}|z^{k+1} - z^k|^2_{M_{\tau,\sigma}} + \frac{1}{2}|z^{k+1} - z|^2_{M_{\tau,\sigma}} - \frac{1}{2}|z^k - z|^2_{M_{\tau,\sigma}} - \left\langle K^* y^{k+1}, x \right\rangle + \left\langle K x^{k+1}, y \right\rangle \\ + g(x^{k+1}) + f^*(y^{k+1}) \le g(x) + f^*(y)$$

PDHG: rate

Therefore, introducing the Lagrangian $\mathcal{L}(x,y) = g(x) - f^*(y) + \langle Kx, y\rangle$ and using:

$$\mathcal{L}(x^{k+1},y) - \mathcal{L}(x,y^{k+1}) = g(x^{k+1}) + \left\langle y, Kx^{k+1}\right\rangle - f^*(y) - g(x) - \left\langle y^{k+1}, Kx\right\rangle + f^*(y^{k+1})$$

we obtain for any $z = (x,y)^T$:

$$\mathcal{L}(x^{k+1},y) - \mathcal{L}(x,y^{k+1}) + \tfrac{1}{2}|z^{k+1} - z^k|^2_{M_{\tau,\sigma}} + \tfrac{1}{2}|z^{k+1} - z|^2_{M_{\tau,\sigma}} \le \tfrac{1}{2}|z^k - z|^2_{M_{\tau,\sigma}}.$$

so that, if $M_{\tau,\sigma} \ge 0$, for any $N \ge 1$,

$$\sum_{k=0}^{N-1} \mathcal{L}(x^{k+1},y) - \mathcal{L}(x,y^{k+1}) + \frac{1}{2}|z^N - z|^2_{M_{\tau,\sigma}} \le \frac{1}{2}|z^0 - z|^2_{M_{\tau,\sigma}}.$$

PDHG: rate

By convexity, we obtain, denoting $Z^N = (X^N, Y^N)^T := \frac{1}{N}\sum_{k=1}^N z^k$:

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{1}{2N}|z^0 - z|^2_{M_{\tau,\sigma}}.$$

If the domains of x, y are bounded we deduce:

$$\mathcal{P}(X^N) - \mathcal{D}(Y^N) \leq \frac{1}{N}\left(\frac{D_x^2}{\tau} + \frac{D_y^2}{\sigma}\right)$$

where $D_\bullet$ are the diameters of the corresponding sets. (Similar to Nemirovsky, 2010, for an extragradient variant.)

PDHG: rate

By convexity, we obtain, denoting $Z^N = (X^N, Y^N)^T := \frac{1}{N}\sum_{k=1}^N z^k$:

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \leq \frac{1}{2N}|z^0 - z|^2_{M_{\tau,\sigma}}.$$

If the domains of x, y are bounded we deduce:

$$\mathcal{P}(X^N) - \mathcal{D}(Y^N) \leq \frac{1}{N}\left(\frac{D_x^2}{\tau} + \frac{D_y^2}{\sigma}\right)$$

where $D_\bullet$ are the diameters of the corresponding sets. (Similar to Nemirovsky, 2010, for an extragradient variant.)

Remark: we just used $\tau\sigma\|K\|^2 \leq 1$ (not $<$). If g, f^* provide additional information on the coerciveness of g, f^* it is enough (in finite dimension) to show convergence of the algorithm.

PDHG: Extensions

- one can over-relax;
- one can ad an "explicit" (co-coercive) term:

we obtain an extension due to L. Condat (in a generalized form to B.C. Vu, referred usually as Condat-Vu's primal-dual algorithm). If h is a convex function with L_h-Lipschitz gradient one writes:

$$\begin{pmatrix} \frac{1}{\tau} I & -K^* \\ -K & \frac{1}{\sigma} I \end{pmatrix} \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix} + \begin{pmatrix} \partial g(x^{k+1}) \\ \partial f^*(y^{k+1}) \end{pmatrix} + \begin{pmatrix} 0 & K^* \\ -K & 0 \end{pmatrix} \begin{pmatrix} x^{k+1} \\ y^{k+1} \end{pmatrix} \ni \begin{pmatrix} -\nabla h(x^k) \\ 0 \end{pmatrix}.$$

Then, this is exactly a foward-backward splitting for two operators and we know that it will converge provided, in the metric $M_{\tau,\sigma}$:

$$C = M_{\tau,\sigma}^{-1} \begin{pmatrix} -\nabla h(x) \\ 0 \end{pmatrix}$$

is μ-co-coercive for some $\mu > 1/2$.

Condat-Vu's variant

That is, if for all z, z':

$$\langle M_{\tau,\sigma}(z - z'), Cz - Cz' \rangle \geq \mu |Cz - Cz'|^2_{M_{\tau,\sigma}}.$$

This can be rewritten

$$\langle x - x', \nabla h(x) - \nabla h(x') \rangle \geq \tau\mu |\nabla h(x) - \nabla h(x')|^2$$

and one will be able to find $\mu > 1/2$ such that this holds as soon as $\tau < 2/L_h$, using Baillon-Haddad's theorem.

Condat-Vu's variant

In the end the method reads:

Input: initial pair of primal and dual points (x^0, y^0), steps $\tau, \sigma > 0$.
for all $k \geq 0$ **do**
find (x^{k+1}, y^{k+1}) by solving

$$x^{k+1} = \mathrm{prox}_{\tau g}(x^k - \tau(K^* y^k + \nabla h(x^k))) \quad (2)$$

$$y^{k+1} = \mathrm{prox}_{\sigma f^*}(y^k + \sigma K(2x^{k+1} - x^k)). \quad (3)$$

end for

which will converge to a fixed point (if it exists) if $\tau < 2/L_h$ and $\tau\sigma\|K\|^2 < 1$. [A rate can also be shown with a proof similar to the previous.]

Condat-Vu's variant

In the end the method reads:

Input: initial pair of primal and dual points (x^0, y^0), steps $\tau, \sigma > 0$.

for all $k \geq 0$ **do**

find (x^{k+1}, y^{k+1}) by solving

$$x^{k+1} = \text{prox}_{\tau g}(x^k - \tau(K^* y^k + \nabla h(x^k))) \quad (2)$$

$$y^{k+1} = \text{prox}_{\sigma f^*}(y^k + \sigma K(2x^{k+1} - x^k)). \quad (3)$$

end for

which will converge to a fixed point (if it exists) if $\tau < 2/L_h$ and $\tau\sigma\|K\|^2 < 1$. [A rate can also be shown with a proof similar to the previous.]

(!) One should additionally check that a fixed point of these iterations solves:

$$\min_x f(Kx) + g(x) + h(x) = \min_x \sup_y \langle y, Kx \rangle - f^*(y) + g(x) + h(x).$$

PDHG: acceleration

The previous method can be accelerated if g or f^* is strongly convex (and even further if **both** are strongly convex), similarly to the forward-backward splitting. We explain how it works, for instance if g is strongly convex. To make the computation a little bit easier we rather write the method as:

$$\begin{aligned} y^{k+1} &= (I + \sigma\partial f^*)^{-1}(y^k + \sigma K(x^k + \theta(x^k - x^{k-1})) \\ x^{k+1} &= (I + \tau\partial g)^{-1}(x^k - \tau K^* y^{k+1}) \,. \end{aligned}$$

for some $\sigma, \tau > 0$, and some $\theta \in [0, 1]$ (we had $\theta = 1$ in the previous parts).

PDHG: acceleration

Actually, the general form considers "old points" $(\bar{x}, \tilde{x}, \bar{y}, \tilde{y})$ and finds a "new point" $(\hat{x}, \hat{y})$ by solving:

$$\hat{y} = (I + \sigma\partial f^*)^{-1}(\bar{y} + \sigma K\tilde{x})$$
$$\hat{x} = (I + \tau\partial g)^{-1}(\bar{x} - \tau K^*\tilde{y}).$$

In particular, if g is μ_g-convex and/or f^* is μ_{f^*}-convex, then for all x, y, one has:

$$g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{2\tau}|x - \bar{x}|^2 \geq g(\hat{x}) + \langle K\hat{x}, \tilde{y}\rangle + \frac{1}{2\tau}|\hat{x} - \bar{x}|^2 + \frac{1 + \tau\mu_g}{2\tau}|x - \hat{x}|^2$$

$$f^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{2\sigma}|y - \bar{y}|^2 \geq f^*(\hat{y}) - \langle K\tilde{x}, \hat{y}\rangle + \frac{1}{2\sigma}|\hat{y} - \bar{y}|^2 + \frac{1 + \sigma\mu_{f^*}}{2\sigma}|y - \hat{y}|^2$$

PDHG: acceleration

$$g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{2\tau}|x-\bar{x}|^2 \geq g(\hat{x}) + \langle K\hat{x}, \tilde{y}\rangle + \frac{1}{2\tau}|\hat{x}-\bar{x}|^2 + \frac{1+\tau\mu_g}{2\tau}|x-\hat{x}|^2$$

$$f^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{2\sigma}|y-\bar{y}|^2 \geq f^*(\hat{y}) - \langle K\tilde{x}, \hat{y}\rangle + \frac{1}{2\sigma}|\hat{y}-\bar{y}|^2 + \frac{1+\sigma\mu_{f^*}}{2\sigma}|y-\hat{y}|^2$$

as before we sum and see that:

$$\begin{aligned}\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) + \frac{1}{2\tau}|\hat{x}-\bar{x}|^2 + \frac{1+\tau\mu_g}{2\tau}|x-\hat{x}|^2 + \frac{1}{2\sigma}|\hat{y}-\bar{y}|^2 + \frac{1+\sigma\mu_{f^*}}{2\sigma}|y-\hat{y}|^2 \\ \leq \frac{1}{2\tau}|x-\bar{x}|^2 + \frac{1}{2\sigma}|y-\bar{y}|^2 \\ + \langle K\hat{x}, y\rangle - \langle Kx, \hat{y}\rangle + \langle K(x-\hat{x}), \tilde{y}\rangle - \langle K\tilde{x}, y-\hat{y}\rangle\,.\end{aligned}$$

Then, we add and remove $\langle K\hat{x}, \hat{y}\rangle$ to rewrite the last terms:

$$\langle K(x-\hat{x}), \tilde{y}-\hat{y}\rangle - \langle K(\tilde{x}-\hat{x}), y-\hat{y}\rangle\,.$$

PDHG: acceleration

$$g(x) + \langle Kx, \tilde{y}\rangle + \frac{1}{2\tau}|x-\bar{x}|^2 \geq g(\hat{x}) + \langle K\hat{x}, \tilde{y}\rangle + \frac{1}{2\tau}|\hat{x}-\bar{x}|^2 + \frac{1+\tau\mu_g}{2\tau}|x-\hat{x}|^2$$

$$f^*(y) - \langle K\tilde{x}, y\rangle + \frac{1}{2\sigma}|y-\bar{y}|^2 \geq f^*(\hat{y}) - \langle K\tilde{x}, \hat{y}\rangle + \frac{1}{2\sigma}|\hat{y}-\bar{y}|^2 + \frac{1+\sigma\mu_{f^*}}{2\sigma}|y-\hat{y}|^2$$

as before we sum and see that:

$$\begin{aligned}\mathcal{L}(\hat{x}, y) - \mathcal{L}(x, \hat{y}) + \frac{1}{2\tau}|\hat{x}-\bar{x}|^2 + \frac{1+\tau\mu_g}{2\tau}|x-\hat{x}|^2 + \frac{1}{2\sigma}|\hat{y}-\bar{y}|^2 + \frac{1+\sigma\mu_{f^*}}{2\sigma}|y-\hat{y}|^2 \\ \leq \frac{1}{2\tau}|x-\bar{x}|^2 + \frac{1}{2\sigma}|y-\bar{y}|^2 \\ + \langle K\hat{x}, y\rangle - \langle Kx, \hat{y}\rangle + \langle K(x-\hat{x}), \tilde{y}\rangle - \langle K\tilde{x}, y-\hat{y}\rangle\,.\end{aligned}$$

Then, we add and remove $\langle K\hat{x}, \hat{y}\rangle$ to rewrite the last terms:

$$\langle K(x-\hat{x}), \tilde{y}-\hat{y}\rangle - \langle K(\tilde{x}-\hat{x}), y-\hat{y}\rangle\,.$$

$\rightarrow$ the best would be to take $\tilde{x} = \hat{x}$ and $\tilde{y} = \hat{y}$ to get rid of these terms... (but then it is totally implicit).

Accelerated PDHG

$$\mathcal{L}(\hat{x},y)-\mathcal{L}(x,\hat{y})+\frac{1}{2\tau}|\hat{x}-\bar{x}|^2+\frac{1+\tau\mu_g}{2\tau}|x-\hat{x}|^2+\frac{1}{2\sigma}|\hat{y}-\bar{y}|^2+\frac{1+\sigma\mu_{f^*}}{2\sigma}|y-\hat{y}|^2$$
$$\leq\frac{1}{2\tau}|x-\bar{x}|^2+\frac{1}{2\sigma}|y-\bar{y}|^2+\langle K(x-\hat{x}),\tilde{y}-\hat{y}\rangle-\langle K(\tilde{x}-\hat{x}),y-\hat{y}\rangle\,.$$

reads in our case:

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau}|x^{k+1}-x^k|^2+\frac{1+\tau\mu_g}{2\tau}|x-x^{k+1}|^2+\frac{1}{2\sigma}|y^{k+1}-y^k|^2+\frac{1+\sigma\mu_{f^*}}{2\sigma}|y-y^{k+1}|^2$$
$$\leq\frac{1}{2\tau}|x-x^k|^2+\frac{1}{2\sigma}|y-y^k|^2+\left\langle K(x-x^{k+1}),\tilde{y}-y^{k+1}\right\rangle-\left\langle K(\tilde{x}-x^{k+1}),y-y^{k+1}\right\rangle.$$

and we can specialize in a semi-implicit form: $\tilde{y}=y^{k+1}$ and $\tilde{x}=x^k+\theta(x^k-x^{k-1})$ for some θ choosen later on, so that the last term becomes:

$$-\left\langle K(x^k+\theta(x^k-x^{k-1})-x^{k+1}),y-y^{k+1}\right\rangle=\left\langle K(x^{k+1}-x^k),y-y^{k+1}\right\rangle-\theta\left\langle K(x^k-x^{k-1}),y-y^{k+1}\right\rangle$$

Accelerated PDHG

We end up with:

$$\mathcal{L}(x^{k+1},y) - \mathcal{L}(x,y^{k+1}) + \frac{1}{2\tau}|x^{k+1}-x^k|^2 + \frac{1}{2\sigma}|y^{k+1}-y^k|^2 + \frac{1+\tau\mu_g}{2\tau}|x-x^{k+1}|^2 + \frac{1+\sigma\mu_{f^*}}{2\sigma}|y-y^{k+1}|^2 - \left\langle K(x^{k+1}-x^k), y-y^{k+1}\right\rangle \le \frac{1}{2\tau}|x-x^k|^2 + \frac{1}{2\sigma}|y-y^k|^2 - \theta\left\langle K(x^k-x^{k-1}), y-y^{k+1}\right\rangle$$

Accelerated PDHG

We end up with:

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau}|x^{k+1}-x^k|^2+\frac{1}{2\sigma}|y^{k+1}-y^k|^2$$
$$+\frac{1+\tau\mu_g}{2\tau}|x-x^{k+1}|^2+\frac{1+\sigma\mu_{f^*}}{2\sigma}|y-y^{k+1}|^2-\left\langle K(x^{k+1}-x^k),y-y^{k+1}\right\rangle$$
$$+\theta\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle\le\frac{1}{2\tau}|x-x^k|^2+\frac{1}{2\sigma}|y-y^k|^2-\theta\left\langle K(x^k-x^{k-1}),y-y^k\right\rangle$$

Accelerated PDHG

We end up with:

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau}|x^{k+1}-x^k|^2+\frac{1}{2\sigma}|y^{k+1}-y^k|^2$$
$$+\frac{1+\tau\mu_g}{2\tau}|x-x^{k+1}|^2+\frac{1+\sigma\mu_{f^*}}{2\sigma}|y-y^{k+1}|^2-\left\langle K(x^{k+1}-x^k),y-y^{k+1}\right\rangle$$
$$+\theta\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle\le\frac{1}{2\tau}|x-x^k|^2+\frac{1}{2\sigma}|y-y^k|^2-\theta\left\langle K(x^k-x^{k-1}),y-y^k\right\rangle$$

Provided we can control the cross term $\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle$ with the terms $\frac{1}{2\tau}|x^{k+1}-x^k|^2+\frac{1}{2\sigma}|y^{k+1}-y^k|^2$ we can hope to obtain a rate of convergence, even linear if $\mu_g>0$ and $\mu_{f^*}>0$. Let us consider the more difficult case $\mu_g>0$, $\mu_{f^*}=0$.

Accelerated PDHG

In this case, we assume θ, σ, τ are varying and depend on k and we write:

$$\begin{aligned}\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x, y^{k+1}) + \frac{1}{2\tau_k}|x^{k+1} - x^k|^2 + \frac{1}{2\sigma_k}|y^{k+1} - y^k|^2 + \theta_k \left\langle K(x^k - x^{k-1}), y^k - y^{k+1} \right\rangle \\ + \frac{1 + \tau_k \mu_g}{2\tau_k}|x - x^{k+1}|^2 + \frac{1}{2\sigma_k}|y - y^{k+1}|^2 - \left\langle K(x^{k+1} - x^k), y - y^{k+1} \right\rangle \\ \le \frac{1}{2\tau_k}|x - x^k|^2 + \frac{1}{2\sigma_k}|y - y^k|^2 - \theta_k \left\langle K(x^k - x^{k-1}), y - y^k \right\rangle\end{aligned}$$

Accelerated PDHG

In this case, we assume θ, σ, τ are varying and depend on k and we write:

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau_k}|x^{k+1}-x^k|^2+\frac{1}{2\sigma_k}|y^{k+1}-y^k|^2+\theta_k\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle$$
$$+\frac{\tau_{k+1}(1+\tau_k\mu_g)}{\tau_k}\frac{1}{2\tau_{k+1}}|x-x^{k+1}|^2+\frac{\sigma_{k+1}}{\sigma_k}\frac{1}{2\sigma_{k+1}}|y-y^{k+1}|^2-\frac{\theta_{k+1}}{\theta_{k+1}}\left\langle K(x^{k+1}-x^k),y-y^{k+1}\right\rangle$$
$$\le\frac{1}{2\tau_k}|x-x^k|^2+\frac{1}{2\sigma_k}|y-y^k|^2-\theta_k\left\langle K(x^k-x^{k-1}),y-y^k\right\rangle$$

so that if we can choose

$$\frac{\tau_{k+1}(1+\tau_k\mu_g)}{\tau_k}=\frac{\sigma_{k+1}}{\sigma_k}=\frac{1}{\theta_{k+1}}>1$$

and let $A_k:=\frac{1}{2\tau_k}|x-x^k|^2+\frac{1}{2\sigma_k}|y-y^k|^2-\theta_k\left\langle K(x^k-x^{k-1}),y-y^k\right\rangle$ it reads:

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau_k}|x^{k+1}-x^k|^2+\frac{1}{2\sigma_k}|y^{k+1}-y^k|^2+\theta_k\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle$$
$$+\frac{\sigma_{k+1}}{\sigma_k}A_{k+1}\le A_k$$

Accelerated PDHG

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau_k}|x^{k+1}-x^k|^2+\frac{1}{2\sigma_k}|y^{k+1}-y^k|^2+\theta_k\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle + \frac{\sigma_{k+1}}{\sigma_k}A_{k+1}\le A_k$$

Then, we use that (denoting to simplify $L:=\|K\|$):

$$-\theta_k\left\langle K(x^k-x^{k-1}),y^k-y^{k+1}\right\rangle\le\frac{\theta_k^2L^2\sigma_k}{2}|x^k-x^{k-1}|^2+\frac{1}{2\sigma_k}|y^k-y^{k+1}|^2$$

to arrive at

$$\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1})+\frac{1}{2\tau_k}|x^{k+1}-x^k|^2-\frac{\theta_k^2L^2\sigma_k}{2}|x^k-x^{k-1}|^2+\frac{\sigma_{k+1}}{\sigma_k}A_{k+1}\le A_k$$

or (after multiplication with σ_k):

$$\sigma_k(\mathcal{L}(x^{k+1},y)-\mathcal{L}(x,y^{k+1}))+\frac{\sigma_k}{2\tau_k}|x^{k+1}-x^k|^2-\frac{\theta_k^2L^2\sigma_k^2}{2}|x^k-x^{k-1}|^2+\sigma_{k+1}A_{k+1}\le\sigma_kA_k.$$

Accelerated PDHG

We now sum from $k = 0$ to $N - 1$ the inequality (we use also $\theta_k \sigma_k = \sigma_{k-1}$):

$$\sigma_k(\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x, y^{k+1})) + \frac{\sigma_k}{2\tau_k}|x^{k+1} - x^k|^2 - \frac{L^2\sigma_{k-1}^2}{2}|x^k - x^{k-1}|^2 + \sigma_{k+1}A_{k+1} \le \sigma_k A_k.$$

We let $T_N = \sum_{k=0}^{N-1}\sigma_k$, $X^N = \frac{1}{T_N}\sum_{k=0}^{N-1}\sigma_k x^{k+1}$, $Y^N = \frac{1}{T_N}\sum_{k=0}^{N-1}\sigma_k y^{k+1}$, so that:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) \le \sum_{k=0}^{N-1}\sigma_k(\mathcal{L}(x^{k+1}, y) - \mathcal{L}(x, y^{k+1}))$$

thanks to the convexity of $\mathcal{L}(\cdot, y) - \mathcal{L}(x, \cdot')$ Then we get:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \sum_{k=1}^{N}\frac{\sigma_{k-1}}{2\tau_{k-1}}|x^k - x^{k-1}|^2 - \sum_{k=0}^{N-1}\frac{L^2\sigma_{k-1}^2}{2}|x^k - x^{k-1}|^2 + \sigma_N A_N \le \sigma_0 A_0$$

or, choosing $x^{-1} = x^0$:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \frac{\sigma_{N-1}}{2\tau_{N-1}}|x^N - x^{N-1}|^2 + \sum_{k=1}^{N-1}\left(\frac{\sigma_{k-1}}{\tau_{k-1}}(1 - L^2\tau_{k-1}\sigma_{k-1})\right)\frac{|x^k - x^{k-1}|}{2}^2 + \sigma_N A_N \le \sigma_0 A_0$$

Accelerated PDHG

Hence: we choose in addition $L^2\sigma_k\tau_k \le 1$ (or $=$ in practice) and end up with:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \frac{\sigma_{N-1}}{2\tau_{N-1}}|x^N - x^{N-1}|^2 + \frac{\sigma_N}{2\tau_N}|x^N - x|^2 + \frac{1}{2}|y^N - y|^2$$
$$- \sigma_N\theta_N \left\langle K(x^N - x^{N-1}), y - y^N \right\rangle \le \frac{\sigma_0}{2\tau_0}|x^0 - x|^2 + \frac{1}{2}|y^0 - y|^2$$

(using again $x^{-1} = x^0$ and recalling $A_k := \frac{1}{2\tau_k}|x - x^k|^2 + \frac{1}{2\sigma_k}|y - y^k|^2 - \theta_k \left\langle K(x^k - x^{k-1}), y - y^k \right\rangle$). We end up estimating again:

$$\sigma_N\theta_N \left\langle K(x^N - x^{N-1}), y - y^N \right\rangle \le \frac{L^2\sigma_{N-1}^2}{2}|x^N - x^{N-1}|^2 + \frac{1}{2}|y - y^N|^2$$

and using $L^2\sigma_{N-1}^2 \le \sigma_{N-1}/\tau_{N-1}$ to find:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \frac{\sigma_N}{2\tau_N}|x^N - x|^2 \le \frac{\sigma_0}{2\tau_0}|x^0 - x|^2 + \frac{1}{2}|y^0 - y|^2$$

Accelerated PDHG

Now we specify the parameters... In order to keep $L^2\sigma_k\tau_k = 1$ (to simplify) we keep $\sigma_k\tau_k = \sigma_0\tau_0 = 1/L^2$. Then, we should have:

$$\frac{\tau_{k+1}(1+\tau_k\mu_g)}{\tau_k} = \frac{\sigma_{k+1}}{\sigma_k} = \frac{1}{\theta_{k+1}} > 1$$

and in particular, $\sigma_{k+1} = \sigma_k/\theta_{k+1}$ and $\tau_{k+1} = \theta_{k+1}\tau_k$, so that:

$$\theta_{k+1} = \frac{1}{\sqrt{1+\mu_g\tau_k}}, \quad \tau_{k+1} = \frac{\tau_k}{\sqrt{1+\mu_g\tau_k}}, \quad \sigma_{k+1} = \sigma_k\sqrt{1+\mu_g\tau_k}.$$

In particular $\sigma_{k+1}^2 = \sigma_k^2 + \kappa\sigma_k$ where $\kappa = \mu_g/L^2$ is an (inverse) condition number. One can then show that: if τ_0 is large, then after very few iterations, $\tau_k \leq 1$: we use

$$\mu_g\tau_{k+1} = \frac{\mu_g\tau_k}{\sqrt{1+\mu_g\tau_k}} \leq \sqrt{\mu_g\tau_k} \quad \Rightarrow \ldots$$

Accelerated PDHG: paramters and rate

$$\mu_g\tau_{k+1} \le \sqrt{\mu_g\tau_k} \Rightarrow \log \mu_g\tau_k \le \frac{1}{2^k}\log \mu_g\tau_0$$

so that, for instance, $\mu_g\tau_k \le 2$ as soon as $k \ge \log_2\log_2(\mu_g\tau_0)$ (or $k=0$), which is always very small. Then, for larger ks, one has $\sigma_k = 1/(L^2\tau_k) \ge \kappa/2$ and:

$$\sigma_{k+1}^2 = \sigma_k^2 + \kappa\sigma_k \ge \sigma_k^2 + \alpha\kappa\sigma_k + (1-\alpha)\frac{\kappa^2}{2} = (\sigma_k + \tfrac{\alpha}{2}\kappa)^2 + \left((1-\alpha) - \frac{\alpha^2}{2}\right)\frac{\kappa^2}{2} = (\sigma_k + \tfrac{\alpha}{2}\kappa)^2$$

if we choose $\alpha = \sqrt{3}-1 \approx 0.73$. Then it follows $\sigma_k \gtrsim (.73\kappa/2)k$ and in particular, $T_N \gtrsim (.73\kappa/4)N(N-1)$ (in fact, one can show $\sigma_k \sim (\kappa/2)k$ and $T_N \sim \kappa N^2/4$).

Accelerated PDHG

Getting back to:

$$T_N(\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N)) + \frac{\sigma_N}{2\tau_N}|x^N - x|^2 \leq \frac{\sigma_0}{2\tau_0}|x^0 - x|^2 + \frac{1}{2}|y^0 - y|^2$$

we see that, taking $(x, y) = (x^*, y^*)$ (for which one can show: $\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) \geq \mu_g |X^N - x^*|^2$) one has

$$|x^N - x^*|^2 + |X^N - x^*|^2 \lesssim \frac{CL^2}{\mu_g^2 N^2}$$

and

$$\mathcal{L}(X^N, y) - \mathcal{L}(x, Y^N) = O(\kappa^{-1} N^{-2})(|x - x^0|^2 + |y - y^0|^2)$$

(for $\sigma_0 < \tau_0$).

It shows an improvement over the non-accelerated method provided $N \gtrsim 1/\kappa$.

Accelerated PDHG

A similar (easier) proof shows an accelerated rate in case both functions are strongly convex.