PGMO Lecture: Vision, Learning and Optimization

4. Primal dual methods

Thomas Pock

Institute of Computer Graphics and Vision, TU Graz

February 11, 2020

Overview

PDHG algorithm

Accelerated version

Extensions

Augmented Lagrangian and ADMM

Saddle-point problem

- We again consider problems of the form

$$\min_{x \in \mathcal{X}} f(Kx) + g(x),$$

  where f, g are convex, l.s.c. and 'simple', and $K : \mathcal{X} \to \mathcal{Y}$ is a bounded linear operator.
- Rewriting the problem as a saddle-point problem

$$\min_{x} \max_{y} \mathcal{L}(x, y) := \langle y, Kx \rangle - f^*(y) + g(x)$$

- The most basic algorithm to find a saddle-point dates back to [Arrow, Hurwicz, Uzawa '58]. It alternates a proximal descent in x and a proximal ascent in y

$$\begin{cases} x^{k+1} = \text{prox}_{\tau g}(x^k - \tau K^* y^k), \\ y^{k+1} = \text{prox}_{\sigma f^*}(y^k + \sigma K x^{k+1}). \end{cases}$$

- Convergence requires boundedness of the domain of f^* and $\tau = 1/\sqrt{k}$

Convergence

- A convergent algorithm is obtained by incorporating so-called extra-gradients [Korpelevich '76], [Popov 81]
- Another simple modification is to replace x^{k+1} in the second line by $2x^{k+1} - x^k$ [P. Cremers, Bischof, Chambolle '09], [Esser et al. '10]

Algorithm 1 PDHG.

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

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

$$\begin{cases} x^{k+1} = \text{prox}_{\tau g}(x^k - \tau K^* y^k) \\ y^{k+1} = \text{prox}_{\sigma f^*}(y^k + \sigma K(2x^{k+1} - x^k)). \end{cases}$$

end for

Relations to the proximal point algorithm

- It can be shown [He, You, Yuan '14] that the algorithm is just an instance of the proximal point algorithm in a certain metric M

$$\begin{pmatrix} K^* x^{k+1} + \partial g(x^{k+1}) \\ -K y^{k+1} + \partial f^*(y^{k+1}) \end{pmatrix} + M \begin{pmatrix} x^{k+1} - x^k \\ y^{k+1} - y^k \end{pmatrix} \ni 0$$

- It turns out that the correct metric M is given by

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

which is positive definite as soon as

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

A more general class of problems

- Let us consider a slightly more general form:

$$\min_{x \in \mathcal{X}} f(Kx) + g(x) + h(x),$$

  where h is a convex function with L_h Lipschitz continuous gradient. The corresponding Lagrangian is given by

$$\mathcal{L}(x, y) := \langle y, Kx \rangle - f^*(y) + g(x) + h(x)$$

- We consider the following more general form of primal-dual iterations [Condat '13] [Vu '13]:

Algorithm 2 General form of primal–dual iteration.

Input: previous points $(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$, steps $\tau, \sigma > 0$.
Output: new points $(\hat{x}, \hat{y}) = \mathcal{PD}_{\tau,\sigma}(\bar{x}, \bar{y}, \tilde{x}, \tilde{y})$ given by

$$\begin{cases} \hat{x} = \text{prox}_{\tau g}(\bar{x} - \tau(\nabla h(\bar{x}) + K^* \tilde{y})), \\ \hat{y} = \text{prox}_{\sigma f^*}(\bar{y} + \sigma K \tilde{x}). \end{cases}$$

Convergence rate

Choosing as in the PDHG algorithm $\bar{x} = x^k$, $\bar{y} = y^k$, $\tilde{y} = y^k$, $\tilde{x} = 2x^{k+1} - x^k$, we can show the following convergence rate:

Theorem

Let $\tau, \sigma > 0$ *and* $(x^0, y^0) \in \mathcal{X} \times \mathcal{Y}$ *be given, and for* $k \geq 0$ *let*

$$(x^{k+1}, y^{k+1}) = \mathcal{PD}_{\tau,\sigma}(x^k, y^k, 2x^{k+1} - x^k, y^k).$$

Assume $\left(\frac{1}{\tau} - L_h\right)\frac{1}{\sigma} \geq L^2$*. Then, for any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$*, we have*

$$\mathcal{L}(X^k, y) - \mathcal{L}(x, Y^k) \leq \frac{1}{2k}\left\| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \right\|_M^2$$

where $X^k = \frac{1}{k}\sum_{i=1}^k x^i$, $Y^k = \frac{1}{k}\sum_{i=1}^k y^i$*. Moreover, if the step size restriction is strict, then* (x^k, y^k) *converge (weakly in infinite dimension) to a saddle point.*

Remark: Note that the true primal-dual gap $\mathcal{G}(X^k, Y^k)$ can be bounded by taking the supremum on both sides, but this requires additional assumptions on the functions f^*, g, h.

Overview

PDHG algorithm

Accelerated version

Extensions

Augmented Lagrangian and ADMM

Acceleration

- Similar to Nesterov's accelerated gradient descent, we can accelerate the primal-dual algorithm by choosing dynamic step size parameters τ_k and σ_k
- This idea has been first proposed in [Zhu, Chan '07] as a heuristic to accelerate the convergence of the AHU algorithm in case of the ROF model
- In contrast to Nesterov's algorithm, who exploits the smoothness of the function, we exploit the strong convexity of either $g + h$ (or f^*)

Algorithm 3 Accelerated primal–dual algorithm 1.

Choose $\tau_0 = 1/(2L_h)$ and $\sigma_0 = L_h/L^2$ (or any τ_0, σ_0 with $\tau_0\sigma_0 L^2 \leq 1$ if $L_h = 0$), $\theta_0 = 0$ and $x^{-1} = x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$,

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

$(x^{k+1}, y^{k+1}) = \mathcal{PD}_{\tau_k, \sigma_k}(x^k, y^k, x^k + \theta_k(x^k - x^{k-1}), y^{k+1})$,

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

end for

Convergence rate

Theorem

Let $(x^k, y^k)_{k\geq 0}$ *be the iterations of the accelerated primal dual algorithm 1. For each* $k \geq 1$, *define* $t_k = \sigma_{k-1}/\sigma_0$, $T_k = \sum_{i=1}^k t_i$ *and the averaged points*

$$(X^k, Y^k) = \frac{1}{T_k} \sum_{i=1}^{k} t_i (x^i, y^i).$$

Then for any $k \geq 1$ *and any* $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$T_k \big(\mathcal{L}(X^k, y) - \mathcal{L}(x, Y^k)\big) \leq \frac{1}{2\tau_0} \left\|x^0 - x\right\|^2 + \frac{1}{2\sigma_0} \left\|y^0 - y\right\|^2.$$

One can show that $1/T_k = O(1/k^2)$. The global gap converges with this rate with additional assumptions on f, for instance that f has full domain.

Complete strongly convex

- In case both $g + h$ and f^* are strongly convex, one can devise another variant with optimal constant step size parameters which yields an optimal linear convergence rate.

Algorithm 4 Accelerated primal–dual algorithm 2.

Choose $x^{-1} = x^0 \in \mathcal{X}$, $y^0 \in \mathcal{Y}$, and $\tau, \sigma, \theta > 0$ satisfying $\theta^{-1} = 1 + \mu_g \tau = 1 + \mu_{f^*} \sigma$ and $\theta L^2 \sigma \tau \leq 1 - L_h \tau$.

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

  $(x^{k+1}, y^{k+1}) = \mathcal{PD}_{\tau,\sigma}(x^k, y^k, x^k + \theta(x^k - x^{k-1}), y^{k+1})$,

end for

Convergence rate

Theorem

Let $(x^k, y^k)_{k\geq 0}$ be the iterations of the accelerated primal-dual algorithm 2. For each $k \geq 1$, define $t_k = \sigma_{k-1}/\sigma_0$, $T_k = \sum_{i=1}^{k} \theta^{-i+1}$ and the averaged points

$$(X^k, Y^k) = \frac{1}{T_k} \sum_{i=1}^{k} \theta^{-i+1} (x^i, y^i).$$

Then, for any $k \geq 1$ and any $(x, y) \in \mathcal{X} \times \mathcal{Y}$,

$$\mathcal{L}(X^k, y) - \mathcal{L}(x, Y^k) \leq \frac{1}{T_k} \left(\frac{1}{2\tau} \left\| x^0 - x \right\|^2 + \frac{1}{2\sigma} \left\| y^0 - y \right\|^2 \right).$$

Observe that $1/T_k = O(\theta^k)$, so this is indeed a linear convergence rate.

Overview

Non-linear proximal terms

- The PDHG algorithm can also be implemented using Bregman distance functions.
- For this, we choose two norms $\|\cdot\|_x$ and $\|\cdot\|_y$ and corresponding Bregman distance functions $D_x(x, \bar{x})$ and $D_y(y, \bar{y})$ which are 1-strongly convex with respect to the norms, that is
$$D_x(x, \bar{x}) \geq \frac{1}{2} \|x - \bar{x}\|_x^2, \quad D_y(y, \bar{y}) \geq \frac{1}{2} \|y - \bar{y}\|_y^2.$$
- The general convergence rates remains the same.
- It might be beneficial if the respective operator norms are smaller.

α-preconditioning

On can avoid the computation of L via replacing τ, σ by preconditioning matrices:

$$M = \begin{pmatrix} T^{-1} & -K^* \\ -K & \Sigma^{-1} \end{pmatrix} \geq 0 \Leftrightarrow \|\Sigma^{\frac{1}{2}} K T^{\frac{1}{2}}\| \leq 1$$

Lemma

Let $T = \mathrm{diag}(\tau_1, ...\tau_n)$ *and* $\Sigma = \mathrm{diag}(\sigma_1, ..., \sigma_m)$.

$$\tau_j = \frac{1}{\sum_{i=1}^m |K_{i,j}|^{2-\alpha}} \,, \quad \sigma_i = \frac{1}{\sum_{j=1}^n |K_{i,j}|^{\alpha}}$$

then for any $\alpha \in [0, 2]$

$$\|\Sigma^{\frac{1}{2}} K T^{\frac{1}{2}}\|^2 = \sup_{x \in X,\, x \neq 0} \frac{\|\Sigma^{\frac{1}{2}} K T^{\frac{1}{2}} x\|^2}{\|x\|^2} \leq 1 \,.$$

The parameter α can be used to vary between pure primal ($\alpha = 0$) and pure dual ($\alpha = 2$) preconditioning

Backtracking linesearch

- If the operator norm $L = \|K\|$ is unknown, one can also implement a backtracking linesearch procedure, preserving all the convergence guarantees and rates [Malitsky, P. '18].

Algorithm 5 PDHG-linesearch.

Input: initial pair of primal and dual points (x^0, y^0), steps $\tau_0 > 0, \mu \in (0,1), \delta \in (0,1), \beta > 0$.
Set $\theta_0 = 1$.
for all $k \geq 1$ **do**
  $x^k = \text{prox}_{\tau_{k-1} g}(x^{k-1} - \tau_{k-1} K^* y^k)$
  Choose any $\tau_k \in [\tau_{k-1}, \tau_{k-1}\sqrt{1+\theta_{k-1}}]$
  loop
    $\theta_k = \tau_k / \tau_{k-1}$, $\bar{x}^k = x^k + \theta_k (x^k - x^{k-1})$,
    $y^{k+1} = \text{prox}_{\beta \tau_k f^*}(y^k + \beta \tau_k K \bar{x}^k)$
    if $\sqrt{\beta}\tau_k \left\|K^* y^{k+1} - K^* y^k\right\| \leq \delta \left\|y^{k+1} - y^k\right\|$ **then**
      break
    else
      $\tau_k = \tau_k \mu$
    end if
  end loop
end for

Discussion

- The parameter β plays the role of the ratio τ/σ, hence the linesearch condition becomes

$$\tau_k \sigma_k \left\| K^* y^{k+1} - K^* y^k \right\|^2 \leq \delta^2 \left\| y^{k+1} - y^k \right\|^2$$

- Using constant step sizes, the algorithm reduces to the standard PDHG algorithm.
- In practice, δ should be close to 1.
- The role of the primal and dual variables should be chosen such that the respective prox $\text{prox}_{\sigma f^*}(\cdot)$ is simpler.
- Note that we can compute $K\bar{x}^k = (1+\theta_k)Kx^k - \theta_k K x^{k-1}$.
- In case $\text{prox}_{\sigma f^*}(\cdot)$ is linear (or affine), no additional matrix-vector products have to be computed.
- For example, if $f^*(y) = \frac{1}{2}\|y - d\|^2$, then $\text{prox}_{\sigma_k f^*}(u) = \frac{u + \sigma_k d}{1+\sigma_k}$ and

$$\begin{aligned} y^{k+1} &= \text{prox}_{\sigma_k f^*}(y^k + \sigma_k K\bar{x}^k) = \frac{y^k + \sigma_k (K\bar{x}^k + d)}{1+\sigma_k}, \\ K^* y^{k+1} &= \frac{1}{1+\sigma_k}(K^* y^k + \sigma_k (K^* K \bar{x}^k + K^* d)). \end{aligned}$$

- Can be extended to situations where the algorithm can be extended and to cases with explicit gradient steps.

Example: ROF model

- Let us recall the ROF model

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|u-d\|^2,$$

- The saddle-point formulation is given by

$$\min_u \max_{\mathbf{p}} \langle \mathrm{D}u, \mathbf{p}\rangle + \frac{1}{2}\|u-d\|^2 - \delta_{\{\|\cdot\|_{2,\infty}\leq\lambda\}}(\mathbf{p}).$$

- The problem is 1-strongly convex in the primal variable, hence we can make use of the accelerated PDHG algorithm using

$$\hat{\mathbf{p}} = \mathrm{proj}_{\{\|\cdot\|_{2,\infty}\leq\lambda\}}(\tilde{\mathbf{p}}) \Leftrightarrow \hat{\mathbf{p}}_{i,j} = \frac{\tilde{\mathbf{p}}_{i,j}}{\max\{1, \frac{1}{\lambda}|\tilde{\mathbf{p}}_{i,j}|_2\}},$$

  and

$$\hat{u} = \mathrm{prox}_{\tau g}(\tilde{u}) \Leftrightarrow \hat{u}_{i,j} = \frac{\tilde{u}_{i,j} + \tau d_{i,j}}{1+\tau}.$$

Practical

`rof-apg-vs-apd.ipynb`

Example: TV-deblurring

- In the next example we consider the image deblurring problem

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|a * u - d\|^2 .$$

- There are 3 possibilities to apply the PDHG algorithm:

(1) Compute proximal map of data term using the FFT

(2) Keep the quadratic term and perform explicit steps

$$\min_u \max_{\mathbf{p}} \langle \mathrm{D}u, \mathbf{p}\rangle + \frac{1}{2}\|Au - d\|^2 - \delta_{\{\|\cdot\|_{2,\infty}\leq\lambda\}}(\mathbf{p}).$$

(3) Additionally dualize the quadratic term

$$\min_u \max_{\mathbf{p},q} \langle \mathrm{D}u, \mathbf{p}\rangle - \delta_{\{\|\cdot\|_{2,\infty}\leq\lambda\}}(\mathbf{p}) + \langle Au, q\rangle - \frac{1}{2}\|q + d\|^2 ,$$

with the proximal map

$$\hat{q} = \mathrm{prox}_{\sigma f_q^*}(\tilde{q}) \Leftrightarrow \hat{q}_{i,j} = \frac{\tilde{q}_{i,j} - \sigma d_{i,j}}{1+\sigma}.$$

Practical

`tv-deconv-pd.ipynb`

Example: TV-ℓ_1 model

- Finally, we consider the completely non-smooth TV-ℓ_1 model, which is given by

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \|u - d\|_1$$

- The saddle-point formulation reads

$$\min_u \max_{\mathbf{p}} \langle \mathrm{D}u, \mathbf{p} \rangle + \|u - d\|_1 - \delta_{\{\|\cdot\|_{2,\infty} \le \lambda\}}(\mathbf{p}).$$

- The PDHG algorithm can be applied with the proximal map $\hat{u} = \text{prox}_{\tau g}(\tilde{u})$ given by

$$\hat{u}_{i,j} = d_{i,j} + \max\{0, |\tilde{u}_{i,j} - d_{i,j}| - \tau\} \cdot \operatorname{sgn}(\tilde{u}_{i,j} - d_{i,j}).$$

Practical

`tv-l1-pd.ipynb`

Overview

PDHG algorithm

Accelerated version

Extensions

Augmented Lagrangian and ADMM

Augmented Lagrangian

- Perhaps one of the oldest and best studied approaches for solving non-smooth convex problems is the "alternating directions methods of multipliers" (ADMM) [Glowinski, Marroco '75], [Gabay, Mercier '76]
- In its standard form, ADMM can be applied to problems of the form

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

- The idea is to introduce a Lagrange multiplier z and write the "augmented Lagrangian" [Hestenes '69], [Powell '69], [Fortin, Glowinski '82]

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

where $\gamma > 0$ is a parameter.

ADMM

- The ADMM algorithm essentially performs a block-coordinate minimization for x, y followed by a gradient ascent on z.

Algorithm 6 ADMM.

Choose $\gamma > 0$, y^0, z^0.

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

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

end for

Relation between ADMM and PDHG

- It turns out that ADMM is equivalent to the PDHG algorithm, if we let

$$\tilde{f}(\xi) := \min_{\{x\,:\,Ax=\xi\}} f(x), \quad \tilde{g}(\eta) := \min_{\{y\,:\,By=\eta\}} g(y),$$

  and apply the PDHG algorithm to the problem

$$\min_{\xi} \max_{z} \langle z, \xi - b \rangle + \tilde{f}(\xi) - \tilde{g}^*(z)$$

- Hence, the complete convergence theory of PDHG can be applied to the ADMM algorithm
- Moreover, we can accelerate the ADMM algorithm in case $\tilde{f}$ or $\tilde{g}^*$ is strongly convex.

Linearized ADMM

- The PDHG algorithm is equivalent to a linearized variant of ADMM which in case $B = I$ is obtained by adding a proximal term to the first line in the ADMM algorithm

$$x^{k+1} = \arg\min_x f(x) - \langle z^k, Ax \rangle + \frac{\gamma}{2} \|b - Ax - y^k\|^2 + \frac{\gamma}{2} \|x - x^k\|_M^2 ,$$

- Choosing the metric M as

$$M = \frac{1}{\lambda} I - A^* A$$

which is positive definite if $\lambda \|A\|^2 \leq 1$.

- Since $z^k = z^{k-1} + \gamma(b - Ax^k - y^k)$ and letting $\sigma = \lambda/\gamma$ that

$$x^{k+1} = \text{prox}_{\sigma f}(x^k + \sigma A^*(2z^k - z^{k-1})),$$

which is exactly the second line of the PDHG algorithm.

Douglas Rachford splitting

- In case $K = I$, the primal-dual algorithm takes the form

$$\begin{cases} x^{k+1} = \text{prox}_{\tau g}(x^k - \tau y^k), \\ y^{k+1} = \text{prox}_{\sigma f^*}(y^k + \sigma(2x^{k+1} - x^k)), \end{cases}$$

  where $\tau\sigma \leq 1$ and hence $\sigma = 1/\tau$.
- Using Moreau's identity and by a change of variables $v^k = x^k - \tau y^k$, we obtain the Douglas-Rachford splitting algorithm [Douglas, Rachford '56], [Lions, Mercier '79]

$$\begin{cases} x^{k+1} = \text{prox}_{\tau g} v^k, \\ v^{k+1} = v^k - x^{k+1} + \text{prox}_{\tau f}(2x^{k+1} - v^k). \end{cases}$$

- Finally, we note that the ADMM is the same as the Douglas-Rachford splitting algorithm, but on the dual formulation of the problem.

Example: TV deblurring using ADMM

- We turn back to the TV deblurring problem

$$\min_u \lambda \|\mathrm{D}u\|_{2,1} + \frac{1}{2}\|Au - d\|^2 = \min_{\mathbf{p}} \lambda \|\mathbf{p}\|_{2,1} + G(\mathbf{p}),$$

  where $\mathbf{p} = (p_1, p_2)$ and

$$G(\mathbf{p}) := \min_{u:\mathrm{D}u=\mathbf{p}} \frac{1}{2}\|Au - d\|^2$$

- The proximal map of G is computed as $\hat{\mathbf{p}} = \mathrm{D}u$, where u solves

$$\min_u \frac{1}{2\tau}\|\mathrm{D}u - \tilde{\mathbf{p}}\|^2 + \frac{1}{2}\|Au - d\|^2 .$$

- The solution is given by

$$\hat{\mathbf{p}} = \mathrm{D}(\mathrm{D}^*\mathrm{D} + \tau A^*A)^{-1}(\mathrm{D}^*\tilde{\mathbf{p}} + \tau A^*d),$$

  which can be efficiently computed using the FFT.
- The proximal map for $\lambda\|\cdot\|_{2,1}$ is given by a standard shrinkage.