An Explicit Primal-Dual Algorithm for Large Non-Differentiable Convex Problems

Ernie Esser (UCI)
Joint work with Xiaoqun Zhang (UCLA) and Tony Chan (HKUST)

4-12-2010
Outline

- A Model Convex Minimization Problem
- A Modification of the Primal Dual Hybrid Gradient (PDHG) Method
- Related Works
- Connection to the Split Inexact Uzawa Method and Convergence
- Application to Convex Problems with Separable Structure
  - Operator Splitting Techniques
  - Constrained Total Variation Deblurring Example
  - Convexified Multiphase Segmentation Example
- Conclusions
A Model Convex Minimization Problem

\[ \min_{u \in \mathbb{R}^m} J(Au) + H(u) \quad (P) \]

\( J, H \) closed proper convex

\( H : \mathbb{R}^m \to (-\infty, \infty] \)

\( J : \mathbb{R}^n \to (-\infty, \infty] \)

\( A \in \mathbb{R}^{n \times m} \)

Assume there exists an optimal solution \( u^* \) to \( (P) \)
The PDHG Method

\[
\min_u J(Au) + H(u) \quad (P)
\]

\[
J(Au) = J^{\ast\ast}(Au) = \sup_p \langle p, Au \rangle - J^*(p)
\]

Saddle Point Formulation:

\[
\min_u \sup_p -J^*(p) + \langle p, Au \rangle + H(u) \quad (PD)
\]

Interpret PDHG as primal-dual proximal point method:

\[
p^{k+1} = \arg\max_{p \in \mathbb{R}^n} -J^*(p) + \langle p, Au^k \rangle - \frac{1}{2\delta_k} \|p - p^k\|_2^2
\]

\[
u^{k+1} = \arg\min_{u \in \mathbb{R}^m} H(u) + \langle A^T p^{k+1}, u \rangle + \frac{1}{2\alpha_k} \|u - u^k\|_2^2
\]

PDHG Pros and Cons

\[ p^{k+1} = \arg \min_{p \in \mathbb{R}^n} J^*(p) - \langle p, Au^k \rangle + \frac{1}{2\delta_k} \| p - p^k \|_2^2 \]

\[ u^{k+1} = \arg \min_{u \in \mathbb{R}^m} H(u) + \langle A^T p^{k+1}, u \rangle + \frac{1}{2\alpha_k} \| u - u^k \|_2^2 \]

Pros:
- Simple iterations
- Explicit. PDHG iterations don’t require solving linear systems involving \( A \), just matrix multiplication by \( A \) and \( A^T \)
- Empirically can be very efficient for well chosen \( \alpha_k \) and \( \delta_k \) parameters

Cons:
- General convergence properties unknown
Modified PDHG (PDHGMp)

PDHGMp:

\[
\begin{align*}
    u^{k+1} &= \arg \min_{u \in \mathbb{R}^m} H(u) + \langle A^T (2p^k - p^{k-1}), u \rangle + \frac{1}{2\alpha} \|u - u^k\|_2^2 \\
    p^{k+1} &= \arg \min_{p \in \mathbb{R}^n} J^*(p) - \langle p, Au^{k+1} \rangle + \frac{1}{2\delta} \|p - p^k\|_2^2
\end{align*}
\]

Related Works:


Split Primal Saddle Point Formulation

Introduce the constraint $w = Au$ in (P) and form the Lagrangian

$$L_P(u, w, p) = J(w) + H(u) + \langle p, Au - w \rangle$$

The corresponding saddle point problem is

$$\max_{p \in \mathbb{R}^n} \inf_{u \in \mathbb{R}^m, w \in \mathbb{R}^n} L_P(u, w, p) \quad (SPP)$$

- If $(u^*, w^*, p^*)$ is a saddle point for (SPP), then $(u^*, p^*)$ is a saddle point for (PD).
- Although $p$ was introduced in $L_P$ as a Lagrange multiplier for the constraint $Au = w$, it has the same interpretation as the dual variable $p$ in (PD).
Consider adding \( \frac{1}{2} \langle u - u^k, (\frac{1}{\alpha} - \delta A^T A)(u - u^k) \rangle \) to the first step of the Alternating Direction Method of Multipliers (ADMM) applied to (SPP), with \( 0 < \alpha < \frac{1}{\delta \|A\|^2} \).

Split Inexact Uzawa applied to (SPP):

\[
\begin{align*}
    u^{k+1} &= \arg \min_{u \in \mathbb{R}^m} H(u) + \langle A^T p^k, u \rangle + \frac{1}{2\alpha} \| u - u^k + \delta \alpha A^T (Au^k - w^k) \|^2_2 \\
    w^{k+1} &= \arg \min_{w \in \mathbb{R}^n} J(w) - \langle p^k, w \rangle + \frac{\delta}{2} \| Au^{k+1} - w \|^2_2 \\
    p^{k+1} &= p^k + \delta (Au^{k+1} - w^{k+1})
\end{align*}
\]

Note: In general we could similarly modify both minimization steps in ADMM, but by only modifying the first step we can obtain an interesting PDHG-like interpretation.

Convergence of SIU on (SPP)

Convergence requires

• Fixed parameters $\alpha > 0$, $\delta > 0$
• $\alpha < \frac{1}{\delta \|A\|^2}$

Then if $u^*$ is optimal for (P) and $w^* = Au^*$,

• $\|Au^k - w^k\|_2 \to 0$
• $J(w^k) \to J(w^*)$
• $H(u^k) \to H(u^*)$
• All convergent subsequences of $(u^k, w^k, p^k)$ converge to a saddle point for (SPP)

Equivalence to Modified PDHG (PDHGMP)

Split Inexact Uzawa applied to (SPP):

\[ u^{k+1} = \arg \min_{u \in \mathbb{R}^m} H(u) + \langle A^T p^k, u \rangle + \frac{1}{2\alpha} \| u - u^k + \delta \alpha A^T (Au^k - w^k) \|^2 \]

\[ w^{k+1} = \arg \min_{w \in \mathbb{R}^n} J(w) - \langle p^k, w \rangle + \frac{\delta}{2} \| Au^{k+1} - w \|^2 \]

\[ p^{k+1} = p^k + \delta (Au^{k+1} - w^{k+1}) \]

Replace \( \delta (Au^k - w^k) \) in the \( u^{k+1} \) update with \( p^k - p^{k-1} \). Combine \( p^{k+1} \) and \( w^{k+1} \) to get

\[ p^{k+1} = (p^k + \delta Au^{k+1}) - \delta \arg \min_{w} J(w) + \frac{\delta}{2} \| w - \frac{(p^k + \delta Au^{k+1})}{\delta} \|^2 \]

and apply Moreau’s decomposition.

PDHGMP:

\[ u^{k+1} = \arg \min_{u \in \mathbb{R}^m} H(u) + \langle A^T (2p^k - p^{k-1}), u \rangle + \frac{1}{2\alpha} \| u - u^k \|^2 \]

\[ p^{k+1} = \arg \min_{p \in \mathbb{R}^n} J^*(p) - \langle p, Au^{k+1} \rangle + \frac{1}{2\delta} \| p - p^k \|^2 \]
Modified PDHG Pros and Cons

Pros:
- As simple as PDHG
- Explicit: avoids solving linear systems in ADMM
- Convergence theory for SIU method applies
- Requires no extra variables beyond original primal and dual variables

Cons:
- Stability restriction may require small step size parameters
- Dynamic step size schemes not theoretically justified
Convex Problems with Separable Structure

Many seemingly more complicated problems can be written in the form (P).

Example:

\[
\sum_{i=1}^{N} \phi_i(B_iA_i u + b_i) + H(u)
\]

\[
= \sum_{i=1}^{N} J_i(A_i u) + H(u)
\]

\[
= J(Au) + H(u),
\]

where \( A = \begin{bmatrix} A_1 \\ \vdots \\ A_N \end{bmatrix} \) and \( J_i(z_i) = \phi_i(B_i z_i + b_i) \). Let \( p = \begin{bmatrix} p_1 \\ \vdots \\ p_N \end{bmatrix} \). Then

\[
J^*(p) = \sum_{i=1}^{N} J^*_i(p_i)
\]
Applying PDHGMP to $\min_u \sum_{i=1}^{N} J_i(A_i u) + H(u)$ yields:

$$u^{k+1} = \arg \min_u H(u) + \frac{1}{2\alpha} \left\| u - \left( u^k - \alpha \sum_{i=1}^{N} A_i^T (2p_i^k - p_i^{k-1}) \right) \right\|^2_2$$

$$p_i^{k+1} = \arg \min_{p_i} J_i^*(p_i) + \frac{1}{2\delta} \left\| p_i - (p_i^k + \delta A_i u^{k+1}) \right\|^2_2 \quad i = 1, \ldots, N$$

- Need $0 < \alpha \delta < \frac{1}{\|A\|^2_2}$ for stability
Application to TV Minimization Problems

Discretize $\|u\|_{TV}$ using forward differences and assuming Neumann BC

$$\|u\|_{TV} = \sum_{r=1}^{M_r} \sum_{c=1}^{M_c} \sqrt{(D^+_c u_{r,c})^2 + (D^+_r u_{r,c})^2}$$

Vectorize $M_r \times M_c$ matrix by stacking columns

Define a discrete gradient matrix $D$ and a norm $\| \cdot \|_E$.

Letting $J = \| \cdot \|_E$ and $A = D$,

$$J(Au) = \|Du\|_E = \|u\|_{TV}$$
Define a directed grid-shaped graph with \( m = M_r M_c \) nodes corresponding to matrix elements \((r, c)\).

3 × 3 example:

For each edge \( \eta \) with endpoint indices \((i, j), i < j\), define:

\[
D_{\eta,k} = \begin{cases} 
-1 & \text{for } k = i, \\
1 & \text{for } k = j, \\
0 & \text{for } k \neq i, j.
\end{cases}
\]

\[
E_{\eta,k} = \begin{cases} 
1 & \text{if } D_{\eta,k} = -1, \\
0 & \text{otherwise}.
\end{cases}
\]

Can use \( E \) to define norm \( \| \cdot \|_E \) on \( \mathbb{R}^e \) by

\[
\|w\|_E = \left\| \sqrt{E^T(w^2)} \right\|_1 = \sum_{i=1}^{m} \left( \sqrt{E^T(w^2)} \right)_i = \sum_{i=1}^{m} \|w_i\|_2
\]

where \( w_i \) is the vector of edge values for directed edges coming out of node \( i \).

For TV regularization, \( J(Au) = \|Du\|_E = \|u\|_{TV} \)
Handling Convex Constraints

For PDHGMp to work well, we want simple, explicit solutions to the minimization subproblems.

Convex constraint $u \in T$ can be handled by adding convex indicator function

$$g_T(u) = \begin{cases} 
0 & \text{if } u \in T \\
\infty & \text{otherwise.}
\end{cases}$$

This leads to a simple update when the orthogonal projection

$$\Pi_T(z) = \arg \min_u g_T(u) + \|u - z\|^2$$

is easy to compute. For example,

$$T = \{z : \|z - f\|_2 \leq \epsilon\} \Rightarrow \Pi_T(z) = f + \frac{z - f}{\max\left(\frac{\|z - f\|}{\epsilon}, 1\right)}$$
Constrained TV Deblurring Example

\[
\min_{\|Ku-f\|_2 \leq \epsilon} \|u\|_{TV}
\]

can be rewritten as

\[
\min_u \|Du\|_E + g_T(Ku),
\]

where \(g_T\) is the indicator function for \(T = \{z : \|z - f\|_2 \leq \epsilon\}\).

In order to treat both \(D\) and \(K\) explicitly, let

\[
H(u) = 0 \quad \text{and} \quad J(Au) = J_1(Du) + J_2(Ku),
\]

where \(A = \begin{bmatrix} D \\ K \end{bmatrix}\).

Write the dual variable as \(p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}\) and apply PDHGMP.
PDHGMp for Constrained TV Deblurring

\[
u^{k+1} = u^k - \alpha_k \left(D^T (2p_1^k - p_1^{k-1}) + K^T (2p_2^k - p_2^{k-1})\right)
\]

\[
p_1^{k+1} = \Pi_X (p_1^k + \delta_k Du^{k+1})
\]

\[
p_2^{k+1} = p_2^k + \delta_k K u^{k+1} - \delta_k \Pi_T \left(\frac{p_2^k}{\delta_k} + Ku^{k+1}\right),
\]

where

\[
X = \{p : \|p\|_{E^*} \leq 1\}
\]

\[
\Pi_X (p) = \frac{p}{E \max \left(\sqrt{E^T (p^2)}, 1\right)}
\]

and where \(\Pi_T\) is again defined by

\[
\Pi_T (z) = f + \frac{z - f}{\max \left(\frac{\|z-f\|_2}{\epsilon}, 1\right)}.
\]
Deblurring Numerical Result

$K$ convolution operator for normalized Gaussian blur with Std. dev. 3

$h$ clean image

$f = Kh + \eta$

$\eta$ zero mean Gaussian noise Std. dev. 1

$\epsilon = 256$

$\alpha = .2, \delta = .55$

Original, blurry/noisy and image recovered from 300 PDHGMp iterations
Other Examples of When PDHGMp is Efficient

- \( J(z) = \|z\|_2 \Rightarrow J^*(p) = g_{\{p: \|p\|_2 \leq 1\}} \)
- \( J(z) = \frac{1}{2\alpha} \|z\|_2^2 \Rightarrow J^*(p) = \frac{\alpha}{2} \|p\|_2^2 \)
- \( J(z) = \|z\|_1 \Rightarrow J^*(p) = g_{\{p: \|p\|_\infty \leq 1\}} \)
- \( J(z) = \|z\|_E \Rightarrow J^*(p) = g_{\{p: \|p\|_{E^*} \leq 1\}} \)
- \( J(z) = \|z\|_\infty \Rightarrow J^*(p) = g_{\{p: \|p\|_1 \leq 1\}} \)
- \( J(z) = \max(z) \Rightarrow J^*(p) = g_{\{p: p \geq 0 \text{ and } \|p\|_1 = 1\}} \)

Note: Although there’s no simple formula for projecting a vector onto the \( l_1 \) unit ball (or its positive face) in \( \mathbb{R}^n \), this can be computed with \( O(n \log n) \) complexity.
Multiphase Segmentation Example

Many other problems deal with same normalization constraint $c \in C$.

Example: Convex relaxation of multiphase segmentation

Goal: Segment a given image, $h \in \mathbb{R}^M$, into $W$ regions where the intensities in the $w^{th}$ region are close to given intensities $z_w \in \mathbb{R}$ and the lengths of the boundaries between regions are not too long.

$$g_C(c) + \sum_{w=1}^{W} \left( \|c_w\|_{TV} + \frac{\lambda}{2} \langle c_w, (h - z_w)^2 \rangle \right)$$

$$C = \{c = (c_1, \ldots, c_W) : c_w \in \mathbb{R}^M, \sum_{w=1}^{W} c_w = 1, c_w \geq 0 \}$$

This is a convex approximation of the related nonconvex functional which additionally requires the labels, $c$, to only take on the values zero and one.

Similar Numerical Approach

Apply PDHGMp:

\[
H(c) = g_C(c) + \frac{\lambda}{2} \left\langle c, \sum_{w=1}^{W} \mathcal{X}_w^T (h - z_w)^2 \right\rangle
\]

\[
J(Ac) = \sum_{w=1}^{W} J_w(D\mathcal{X}_w c),
\]

where \( A = \begin{bmatrix} D\mathcal{X}_1 & \vdots & D\mathcal{X}_W \end{bmatrix} \), \( \mathcal{X}_w c = c_w \) and

\[
J_w(D\mathcal{X}_w c) = \|D\mathcal{X}_w c\|_E = \|Dc_w\|_E = \|c_w\|_{TV}.
\]

PDHGMp iterations:

\[
c^{k+1} = \Pi_C \left( c^k - \alpha \sum_{w=1}^{W} \mathcal{X}_w^T (D^T (2p_w^k - p_w^{k-1}) + \frac{\lambda}{2} (h - z_w)^2) \right)
\]

\[
p_w^{k+1} = \Pi_X \left( p_w^k + \delta D\mathcal{X}_w c^{k+1} \right) \quad \text{for } w = 1, \ldots, W.
\]
Segmentation Numerical Result

\[ \lambda = 0.0025 \quad z = \begin{bmatrix} 75 & 105 & 142 & 178 & 180 \end{bmatrix} \]

\[ \alpha = \delta = \frac{0.995}{\sqrt{40}} \]

Threshold \( c \) when each \( \| c_{w}^{k+1} - c_{w}^{k} \|_{\infty} < 0.01 \) (150 iterations)

Segmentation of Brain Image Into 5 Regions

Modifications: We can also add \( \mu_w \) parameters to regularize differently the lengths of the boundaries of each region and alternately update the averages \( z \) when they are not known beforehand.
Conclusions About Modified PDHG

• Simple, explicit iterations
• Convergence theory for SIU method applies
• Requires few assumptions: essentially just convexity of $J$ and $H$
• Widely applicable for many convex models with separable structure
• Empirically efficient for many interesting large scale convex optimization problems
• Dynamic step size schemes can help but aren’t theoretically justified