Solving DC Programs that Promote Group 1-Sparsity

Ernie Esser
Contains joint work with Xiaoqun Zhang, Yifei Lou and Jack Xin

SIAM Conference on Imaging Science
Hong Kong Baptist University
May 14 2014

A General Convex Problem

$$\min_u \sum_{i=1}^{N_F} F_i(A_i u - b_i) + G(u) \qquad \text{s.t.} \qquad A_i u = b_i, \ \ i = N_F + 1, \dots, N \tag{P}$$

$$F_i : \mathbb{R}^{m_i} \to (-\infty, \infty], G : \mathbb{R}^m \to (-\infty, \infty] \text{ convex}$$

Can include indicator functions for convex sets defined by

$$\iota_C(u) = \begin{cases} 0 & u \in C \\ \infty & \text{otherwise} \end{cases}$$

Algorithms

Many primal dual methods can effectively solve (P) as long as the functions F_i, G or their conjugates F_i^*, G^* have simple proximal mappings or have Lipschitz continuous gradients:

- Alternating Direction Method of Multipliers (ADMM)
- Split Inexact Uzawa [Zhang, Burger, Osher 2010]
- Modified PDHG [Chambolle, Pock 2010], [He, Yuan 2012], [Esser, Zhang, Chan 2010]
- Bregman ADMM [Wang, Banerjee 2013]
- Accelerated Linearized ADMM [Ouyang, Chen, Lan, Pasiliao 2014]
- Bregman Operator Splitting (BOS) [Zhang, Burger, Bresson, Osher 2010]
- BOSVS [Chen, Hager, Yashtini, Ye, Zhang 2012]

What if G is not convex?

Difference of Convex, Majorization/Minimization Strategy

Linearize the nonconvex part and possibly add and subtract a strongly convex term to define

$\tilde{G}(u, u^n) \geq G(u)$ such that $\tilde{G}$ is strongly convex and $\tilde{G}(u^n, u^n) = G(u^n)$

Then solve a sequence of convex problems

$$u^{n+1} = \arg\min_u \sum_{i=1}^{N_F} F_i(A_i u - b_i) + \tilde{G}(u, u^n)$$
$$\text{s.t.} \quad A_i u = b_i, \quad i = N_F + 1, ..., N$$

Nonconvex Penalties to Promote Group 1-Sparsity

An interesting class of applications are problems requiring group 1-sparsity such as

$$\min_x \frac{\mu}{2}\|b - \begin{bmatrix} A_1 & \dots & A_N \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} \|^2 + \dots \quad \text{s.t.} \quad x_i \geq 0, \|x_i\|_0 \leq 1$$

One effective strategy is to replace $\|x_i\|_0 \leq 1$ with $\gamma(\|x_i\|_1 - \|x_i\|_2)$ or $\gamma(\frac{\|x_i\|_1}{\|x_i\|_2})$ and use the DC majorization/minimization strategy.

Details in *A Method for Finding Structured Sparse Solutions to Non-negative Least Squares Problems with Applications*, with Yifei Lou and Jack Xin, 2013.

A Binary Group 1-Sparsity Model

Many applications additionally require the elements of x_i to be 0 or 1. A difference of convex model for this is to constrain each x_i to the unit simplex

$$x \in S := \{x : x \geq 0, \|x_i\|_1 = 1\} \quad i = 1, ..., N$$

and minimize a concave quadratic $-\gamma\|x\|^2$.

Some imaging applications of binary labeling models:

- Image segmentation
- Nonlocal patch-based image inpainting
- Linear unmixing for hyperspectral images
- Point matching for image registration
- Phase Unwrapping

DC versus Direct Application of Convex Optimization Methods

- The difference of convex approach can guarantee limit points are stationary points.
- If the convex subproblem is expensive, solving a sequence of them may cost too much.
- Direct application of convex optimization methods such as modified PDHG to nonconvex problems lacks convergence guarantees in general but works well for some applications.

Modified PDHG

To simplify notation, let $F_i = \iota_{\{0\}}$ for $i = N_F + 1, ..., N$ and consider

$$\min_u \sum_{i=1}^{N} F_i(A_i u - b_i) + G(u)$$

$$F_i(A_i u - b_i) = F_i^{**}(A_i u - b_i) = \sup_{p_i} \langle p_i, A_i u - b_i \rangle - F_i^*(p_i)$$

Find saddle point of $\min_u \sup_p -F^*(p) + \langle p, Au - b \rangle + G(u)$ by iterating

$$u^{k+1} = \arg\min_u G(u) - \langle A^T p^k, u \rangle + \frac{1}{2\alpha} \| u - u^k \|^2_{M^{-1}}$$

$$p^{k+1} = \arg\min_p F^*(p) + \langle p, A(2u^{k+1} - u^k) - b \rangle + \frac{1}{2\delta} \| p - p^k \|^2$$

where $\alpha\delta \leq \frac{1}{\|AMA^T\|}$ (Can also derive via linearized ADMM or SIU)

Modified PDHG for Binary Labeling Problems

Suppose we want u to be 1-sparse and binary.

$$\text{Let } G(u) = \iota_S(u) - \gamma\|u\|^2 \text{ and let } \quad M = \frac{\gamma_0}{2\alpha\gamma}\mathrm{I} \quad \gamma_0 \in (0,1)$$

$$\text{Then} \qquad u^{k+1} = \Pi_S\left((\frac{1}{1-\gamma_0})(u^k - \alpha MA^T p^k) \right)$$

The only change in the nonconvex version is to lengthen $u^k - \alpha MA^T p^k$ slightly before projecting onto the simplex.

This empirically works if γ_0 is small and γ is large. If not, iterates can oscillate.

More algorithm details are in *Nonlocal Patch-Based Imaging Inpainting Through Minimization of a Sparsity Promoting Nonconvex Functional* with Xiaoqun Zhang (preprint).

Nonconvex Proximal Point Interpretation

Let $M = \mathrm{I}$ and denote the normal cone of S by $N_S(u) = \{s : \langle s, v - u\rangle \leq 0 \text{ for all } v \in S\}$. Then

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} \in \begin{pmatrix} \partial F^*(p^{k+1}) + Au^{k+1} - b \\ N_S(u^{k+1}) - A^T p^{k+1} - 2\gamma u^{k+1} \end{pmatrix} + \begin{bmatrix} \frac{1}{\delta}\mathrm{I} & A \\ A^T & \frac{1}{\alpha} \end{bmatrix} \begin{pmatrix} p^{k+1} - p^k \\ u^{k+1} - u^k \end{pmatrix}$$

where $\alpha\delta \leq \frac{1}{\|A^T A\|}$.

See [He, Yuan 2012] for the convex case.

Example: Phase Unwrapping

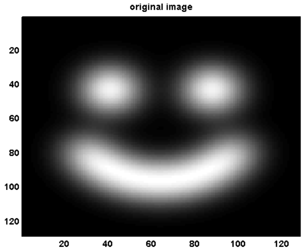
$$d = (h \mod 2\pi) + \text{noise}$$

Assuming h is smooth, recover it from d.

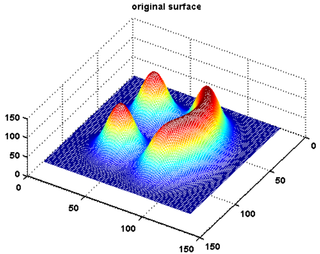
Example application: Interferometric Synthetic Aperture Radar for generating elevation maps (see [Richards 2007])

- Measure wrapped phase at two nearby SAR apertures
- Obtain wrapped phase difference by wrapping difference of wrapped phases
- Use trigonometry to interpret as measurements of relative height modulo $2\pi a$ for some factor a

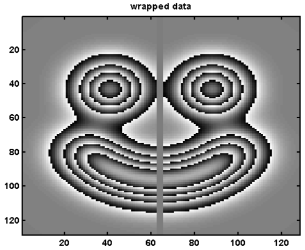
Original image



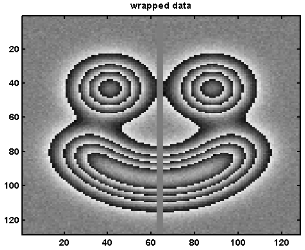
Original image as surface



Wrapped image missing stripe



With noise (SNR = 16.9)



Proposed Model

unknown $u = (x, y, s)$
x - denoised wrapped image
y - unwrapped image
s - 3 labels per edge
X_i - selects ith label from s
$s \geq 0$, $X_1 s + X_2 s + X_3 s = 1$
want $s_{ij} \in \{0, 1\}$

D - discrete gradient
$X_1 s$ - 1 if no jump
$X_2 s$ - 1 if jump by $2\pi a$
$X_3 s$ - 1 if jump by $-2\pi a$
X_d - crops to where data defined
X_D - where differences defined
X_0 - reference no jump region

$$\min_{x,y,s} \lambda \|X_1 s - 1\|_1 + \frac{\beta}{2}\|Dy\|^2 + \iota_S(s) - \gamma\|s\|^2$$

$$\|X_d x - d\|_2 \leq \epsilon$$

$$\|Dy\|_\infty \leq (\pi - \eta)a$$

$$X_D D y = X_D D x + 2\pi a(X_3 s - X_2 s)$$

$$X_0 y = X_0 x$$

Parameters

$$\gamma = 200, \gamma_0 = .002, \beta = .1, \lambda = 100, \alpha\delta < \frac{1}{\|AMA^T\|}$$

$$\text{Let } M = \begin{bmatrix} \mathrm{I} & & \\ & \mathrm{I} & \\ & & M_3 \mathrm{I} \end{bmatrix}, \qquad M_3 = \frac{\gamma_0}{2\alpha\gamma}$$

(For the convex case, let $\gamma = 0$ and define M_3 independently.)

ϵ and η for the constraints $\|X_d x - d\|_2 \leq \epsilon$ and $\|Dy\|_\infty \leq (\pi - \eta)a$ reflect assumptions about the noise and sampling respectively.

Primal Iterates

$$
\begin{aligned}
x^{k+1} &= x^k + \alpha(-X_d^T p_3^k - D^T X_D^T p_4^k - X_0^T p_5^k) \\
y^{k+1} &= y^k + \alpha(-D^T p_2^k + D^T X_D^T p_4^k + X_0^T p_5^k) \\
s^{k+1} &= \Pi_S \left((\frac{1}{1-\gamma_0})(s^k - \alpha M_3 X_1^T p_1^k + 2\pi a \alpha M_3 (X_2^T - X_3^T) p_4^k) \right)
\end{aligned}
$$

Dual Iterates

$$p_1^{k+1} = \Pi_{\|\cdot\|_\infty \leq 1}(p_1^k + \delta X_1(2s^{k+1} - s^k) - \delta)$$

$$p_2^{k+1} = p_2^k + \delta D(2y^{k+1} - y^k) - \delta \Pi_{\|\cdot\|_\infty \leq a(\pi-\eta)} \left(\frac{p_2^k + \delta D(2y^{k+1} - y^k)}{\beta + \delta} \right)$$

Note this combines $\frac{\beta}{2}\|z\|^2 + \iota_{\|\cdot\|_\infty \leq (\pi-\eta)a}(z)$ by observing it is the conjugate of the Huber norm.

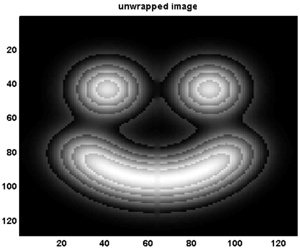
$$p_3^{k+1} = p_3^k + \delta(X_d(2x^{k+1} - x^k) - d) - \Pi_{\|\cdot\|_2 \leq \epsilon\delta}(p_3^k + \delta(X_d(2x^{k+1} - x^k) - d))$$

$$p_4^{k+1} = p_4 - \delta(X_D D(2y^{k+1} - y^k - 2x^{k+1} + x^k) + 2\pi a(X_2 - X_3)(2s^{k+1} - s^k))$$

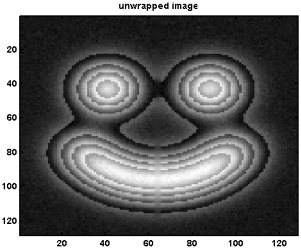
$$p_5^{k+1} = p_5^k - \delta(X_0(2y^{k+1} - y^k - 2x^{k+1} + x^k))$$

Unwrapped Images

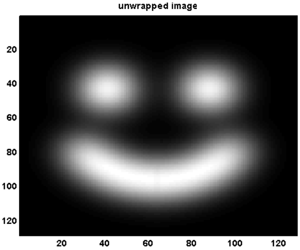
Convex, no noise



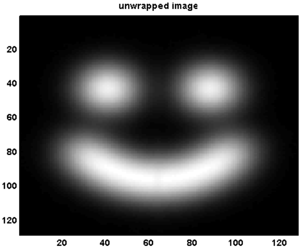
Convex, with noise



Nonconvex, no noise

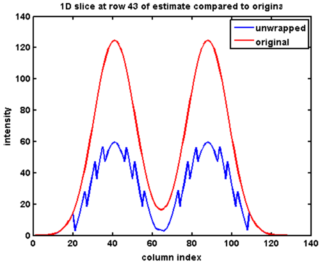


Nonconvex, with noise

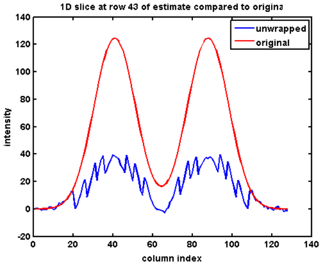


Recovered 1D Slice

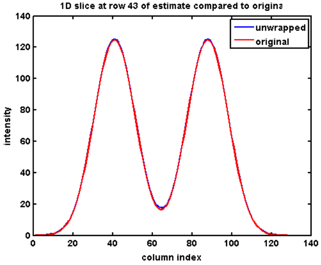
Convex, no noise



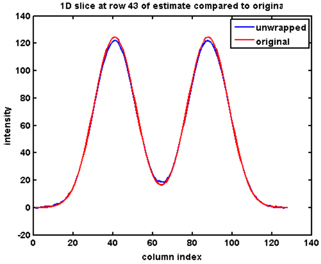
Convex, with noise



Nonconvex, no noise

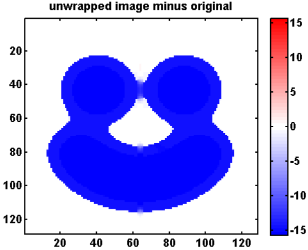


Nonconvex, with noise

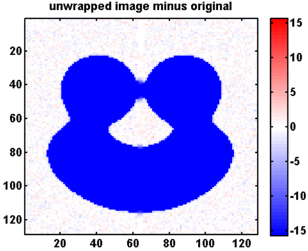


Recovery Error

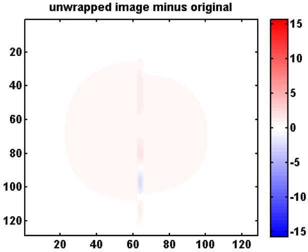
Convex, no noise



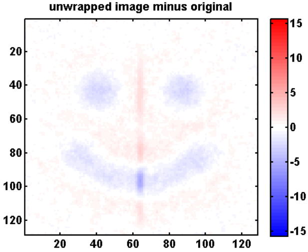
Convex, with noise



Nonconvex, no noise

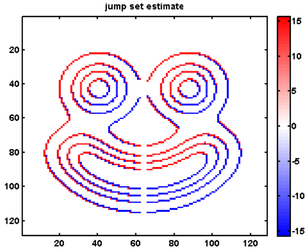


Nonconvex, with noise

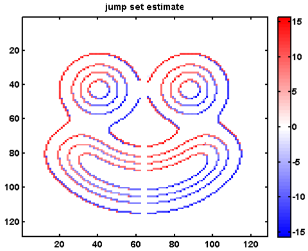


Estimated Jump Set

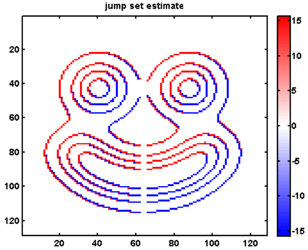
Convex, no noise



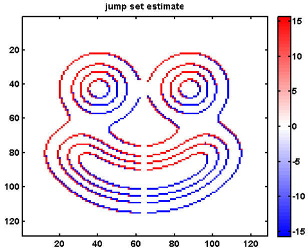
Convex, with noise



Nonconvex, no noise

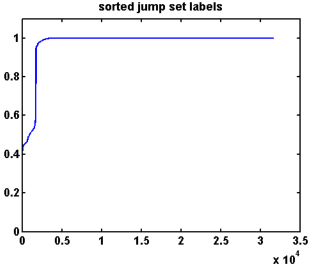


Nonconvex, with noise

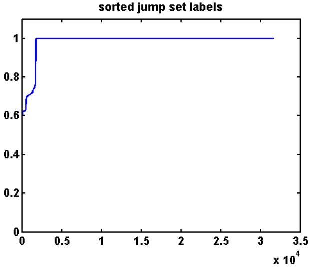


Sorted 'no jump' Labels

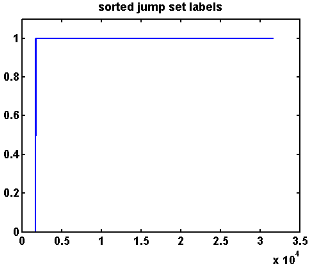
Convex, no noise



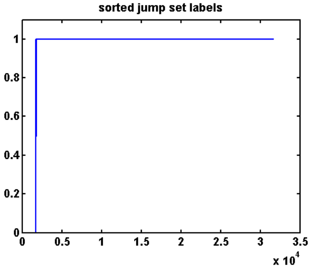
Convex, with noise



Nonconvex, no noise



Nonconvex, with noise



Conclusions and Open Questions

- Modified PDHG can empirically find good local minima for some nonconvex problems where the nonconvexity is of the form of $\iota_S(s) - \gamma\|s\|^2$ (maximizing the l_2 norm of s over a unit simplex).
- An application to 2D phase unwrapping was demonstrated.
- Are there conditions on the parameters that can guarantee at least that limit points are stationary points?