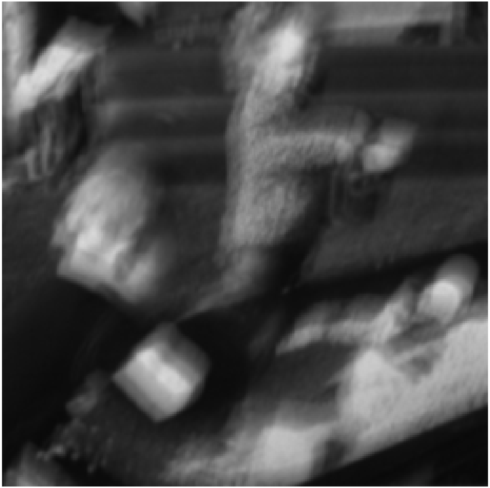
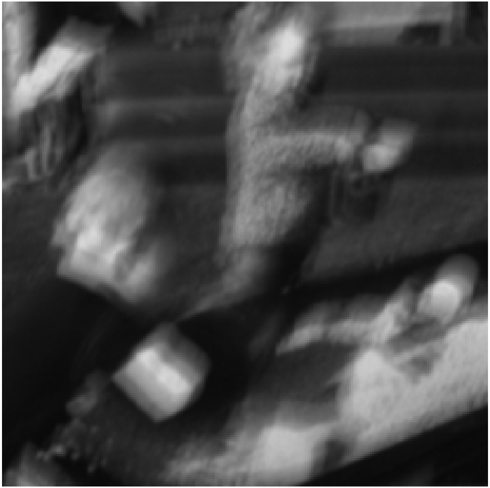
The ASAP algorithm for nonsmooth nonconvex optimization. Applications in imagery

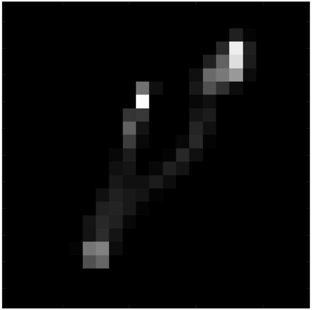
Pauline Tan

Cosmostat Day on Machine Learning in Astrophysics

Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Université
Joint work with Mila Nikolova (CNRS)
and Fabien Pierre (Loria, Université de Lorraine)







Variational approach

w : blurred and noisy image
x : natural sharp image
y : blur kernel in the simplex Σ_M
n : white Gaussian noise

$$w = x \star y + n$$

Joint estimation of the sharp image and the blur kernel

$$\min_{\substack{x \\ y \in \Sigma_M}} J(x, y) = \|x \star y - w\|^2 + \mu \, \mathrm{TV}(x) + \nu \, \|y\|^2 \qquad \mu, \nu > 0$$

TV: total variation regularization
$\star$ is the convolution product

Nonconvex and nonsmooth block-optimisation problems

- joint estimation
- matrix factorization
- parameter estimation
- dimension reduction

Why joint estimation instead of alternate estimation?
robustness, initialization, accuracy

Introduction

Increasing interest in regularized block biconvex (multiconvex) nonconvex optimization problems

$$J(x, y) = F(x) + G(y) + H(x, y)$$

- $(x, y) \in U \times V$ where U and V are finite-dimensional real spaces
- H is block-biconvex: $x \mapsto H(x, y)$ and $y \mapsto H(x, y)$ are convex
- Such $H : U \times V \to \mathbb{R}$ is typically **nonconvex**

Related work

The most intuitive way to solve the problem:
Block Coordinate Descent (BCD) $\rightarrow$ alternating partial minimization

$$\begin{cases} x^{k+1} \in \arg\min_x J(x, y^k) \\ y^{k+1} \in \arg\min_y J(x^{k+1}, y) \end{cases}$$

Introduced by Hildreth in 1957

Proximal regularization of BCD $\rightarrow$ alternating implicit gradient descent

$$\begin{cases} x^{k+1} \in \arg\min_x J(x, y^k) + \frac{1}{2\tau}\|x - x^k\|^2 & = \text{prox}_{\tau J(\cdot, y^k)}(x^k) \\ y^{k+1} \in \arg\min_y J(x^{k+1}, y) + \frac{1}{2\sigma}\|y - y^k\|^2 & = \text{prox}_{\sigma J(x^{k+1}, \cdot)}(y^k) \end{cases}$$

Introduced for convex objectives by Auslander (1992)
Properly extended to some nonconvex objectives by Xu and Yin (2013)

Alternating proximal gradient / forward-backward splitting

In general, the updates for BCD and prox-BCD are not explicit
$\rightarrow$ use a forward-backward splitting

Assume $$J(x,y) = \underbrace{F(x) + G(y)}_{\text{nonsmooth and prox-friendly}} + \underbrace{H(x,y)}_{\text{differentiable}}$$

Replace $$\text{prox}_{\tau J(\cdot, y^k)}(x^k) = \text{prox}_{\tau(F + H(\cdot, y^k))}(x^k)$$

by $$\underbrace{\text{prox}_{\tau^k F}}_{\text{implicit /}}(x^k \underbrace{- \tau^k \nabla_x H(x^k, y^k)}_{\text{explicit gradient descent}})$$ (same for y^k)

Introduced by Xu and Yin (2013)
and Bolte, Sabach and Teboulle (PALM, 2014)
Tremendous improvement compared to previous schemes!

But: Step-sizes (τ^k, σ^k) depend on the Lipschitz constants of $\nabla_x H(\cdot, y^k)$ and $\nabla_y H(x^{k+1}, \cdot)$ $\rightarrow$ estimated at each update

Motivation and proposed ASAP algorithm

Choices for H in applications

$$H(x, y) = \|L(x \cdot y) - w\|^2$$

L linear, w data matrix, and $\cdot$ a product (Hadamard, matrix, scalar, etc.)

Can we avoid that F, G are simple in order to address a wider class of applications?
Can we circumvent the need to update the step-sizes at each iteration?

$\rightarrow$ Use the strong structural property of H
$\rightarrow$ F, G can be smooth:

Many methods with (F, G) nonsmooth use smoothed versions in the algorithms
Smoothed nonsmooth functions are customarily used for sparse recovery

Optimization model

3-block function

$$J(x, y) = F(x) + G(y) + H(x, y)$$

Assumption (H1)

(a) $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ is lowerbounded

(b) $F : U \to \mathbb{R}$ and $G : V \to \mathbb{R}$ are continuously differentiable
∇F is Lipschitz continuous with constant $L_{\nabla F}$
∇G is Lipschitz continuous with constant $L_{\nabla G}$

(c) $H : U \times V \to \mathbb{R} \cup \{+\infty\}$ is proper and lowerbounded

Proposed ASAP algorithm

Simple **Alternating Structure-Adapted Proximal** gradient descent (ASAP) algorithm:

Set $0 < \tau < 1/L_{\nabla F}$ and $0 < \sigma < 1/L_{\nabla G}$
For each $k = 0, 1, 2, \dots,$ compute

$$\begin{cases} x^{k+1} \in \text{prox}_{\tau H(\cdot, y^k)}(x^k - \tau \nabla F(x^k)) \\ y^{k+1} \in \text{prox}_{\sigma H(x^{k+1}, \cdot)}(y^k - \sigma \nabla G(y^k)) \end{cases}$$

$\rightarrow$ Initialisation in dom$J$
$\rightarrow$ Iterates (x^k, y^k) uniquely defined if H is biconvex (even if F, G are nonconvex) and stepsizes twice larger
$\rightarrow$ $J(x^k, y^k) < +\infty$ for any $k \geq 1$
$\rightarrow$ The step-sizes are computed once

(Proximity operator) Let h be a proper function.

$$\text{prox}_h(u) := \arg\min_x \left\{ h(x) + \frac{1}{2} \|x - u\|^2 \right\}$$

ASAP versus PALM

ASAP: For each $k = 0, 1, 2, \ldots,$ compute

$$\begin{cases} x^{k+1} \in \operatorname{prox}_{\tau H(\cdot, y^k)}(x^k - \tau \nabla F(x^k)) \\ y^{k+1} \in \operatorname{prox}_{\sigma H(x^{k+1}, \cdot)}(y^k - \sigma \nabla G(y^k)) \end{cases}$$

τ depends on $L_{\nabla F}$, σ depends on $L_{\nabla G}$

PALM: For each $k = 0, 1, 2, \ldots,$ compute

$$\begin{cases} x^{k+1} \in \operatorname{prox}_{\tau_k F}(x^k - \tau_k \nabla_x H(x^k, y^k)) \\ y^{k+1} \in \operatorname{prox}_{\sigma_k G}(y^k - \sigma_k \nabla_y H(x^{k+1}, y^k)) \end{cases}$$

τ_k depends on $L_{\nabla_x H(\cdot, y^k)}$, σ_k depends on $L_{\nabla_y H(x^{k+1}, \cdot)}$

A generic example of objective functions

$$J(x,y) :=$$

$$\underbrace{\sum_i f_i(\|A_i x\|)}_{=F(x)} + \underbrace{\sum_j g_j(\|B_j y\|)}_{=G(y)} + \underbrace{h(\|b(x,y) - w\|) + \chi_{\mathcal{D}_x}(x) + \chi_{\mathcal{D}_y}(y)}_{=H(x,y)}$$

A_i, B_j linear mappings, b bilinear mapping (e.g. $b(x,y) = x * y$)
Common choices for f_i, g_j, $h : \mathbb{R} \to \mathbb{R}$ ($\alpha > 0$):

(a) $|t|^p$, $p > 1$
(b) $(t^2 + \alpha)^{p/2}$, $0 < p \leq 1$
(c) $|t| - \alpha \log(1 + |t|/\alpha)$
(d) $t^2/(\alpha + t^2)$
(e) $\log(1 + t^2/\alpha)$
(f) with $0 < p \leq 1$
$$\begin{cases} (t - \alpha/2)^p & \text{if } |t| > \alpha \\ (t^2/(2\alpha))^p & \text{if } |t| \leq \alpha \end{cases}$$

(b)-(f) for $\alpha \searrow 0$ provide good smooth approximations of nonsmooth functions.
(b) and (f) for $0 < p < 1$ are smooth approximations of ℓ_p, $0 < p < 1$.

Convergence in value

Let $\rho_x := \frac{1}{\tau} - \frac{L_{\nabla F}}{2} > 0$, $\rho_y := \frac{1}{\sigma} - \frac{L_{\nabla G}}{2} > 0$, $\rho := \min\{\rho_x, \rho_y\}$.

Proposition 1 Let (x^k, y^k) be generated by ASAP under (H1).
(a) (sufficient decrease)

$$\begin{aligned} J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^{k-1}) + \rho_x \|x^{k-1} - x^k\|^2 \\ J(x^k, y^{k-1}) &\geq J(x^k, y^k) + \rho_y \|y^{k-1} - y^k\|^2 \\ J(x^{k-1}, y^{k-1}) &\geq J(x^k, y^k) + \rho \left(\|x^{k-1} - x^k\|^2 + \|y^{k-1} - y^k\|^2\right) \end{aligned}$$

(b) (convergence in value) The sequences $J(x^k, y^k)$ and $J(x^k, y^{k-1})$ converge to the same value J^*

(c) $\sum_{k=1}^{+\infty} (\|x^{k-1} - x^k\|^2 + \|y^{k-1} - y^k\|^2) < +\infty$,

hence $\|x^{k-1} - x^k\| \to 0$ and $\|y^{k-1} - y^k\| \to 0$

Additional assumptions

3-block function

$$J(x, y) = F(x) + G(y) + H(x, y)$$

Assumption (H2)

(a) $J : U \times V \to \mathbb{R} \cup \{+\infty\}$ is lowerbounded

(b) $F : U \to \mathbb{R}$ and $G : V \to \mathbb{R}$ are continuously differentiable
∇F is Lipschitz continuous with constant $L_{\nabla F}$
∇G is Lipschitz continuous with constant $L_{\nabla G}$

(c) $H : U \times V \to \mathbb{R} \cup \{+\infty\}$ is proper and lowerbounded such that
(c-1) H is continuous on its closed domain
(c-2) $\partial_x H(x, y) \times \partial_y H(x, y) \subset \partial H(x, y)$
(c-3-i) $H(x, y) = h(x, y) + f(x) + g(y)$ with h continuous, f, g continuous on their resp. domain, $x \mapsto h(x, y)$ differentiable and $y \mapsto \nabla_x h(x, y)$ locally Lipschitz OR $\nabla_x h$ continuous
(c-3-ii) OR H is biconvex

Approaching the set of critical points

Boundedness assumption The set of limit points of (x^k, y^k)

$$\mathcal{L}(x^0, y^0) := \{(x^*, y^*) \mid \exists k_j \text{ s.t. } (x^{k_j}, y^{k_j}) \to (x^*, y^*)\}$$

is nonempty and bounded (e.g. $\text{dom} J$ is bounded).

Proposition 2 Let (x^k, y^k) be generated by ASAP under (H2).
Let $\mathcal{C} := \{(x, y) \mid 0 \in \partial J(x, y)\}$ set of critical points of J.

(a) $\mathcal{L}(x^0, y^0) \subset \mathcal{C}$

(b) $\text{dist}\big((x^k, y^k), \mathcal{C}\big) \to 0$

Subgradient convergence

Proposition 3 (subgradient convergence) Let (x^k, y^k) be generated by ASAP under (H2)-(c-3-i) with $y \mapsto \nabla_x H(x, y)$ locally Lipschitz. Then, there exists $\beta > 0$ and $(q_x^k, q_y^k) \in \partial J(x^k, y^k)$ s.t.

$$\|(q_x^k, q_y^k)\| \leq \beta \|(x^{k-1} - x^k, y^{k-1} - y^k)\|$$

with $\beta := \max\left\{\sqrt{2}\left(L_{\nabla F} + \frac{1}{\tau}\right), \sqrt{\left(L_{\nabla G} + \frac{1}{\sigma}\right)^2 + 2\xi^2}\right\}$

Kurdyka-Łojasiewicz property

KŁ property A proper l.s.c. function f has the Kurdyka-Łojasiewicz (KŁ) property at $x^* \in \mathrm{dom}\partial f$ if there exists $\eta \in (0, +\infty]$, a neighborhood $\mathcal{O}(x^*)$ of x^*, $\theta \in [0, 1)$, and a constant $\kappa > 0$ s.t.

$$\forall x \in \widetilde{\mathcal{O}}(x^*), \quad \kappa\,\mathrm{dist}(0, \partial f(x)) \geq |f(x) - f(x^*)|^{\theta},$$

with $\widetilde{\mathcal{O}}(x^*) := \mathcal{O}(x^*) \cap \left\{ x \in U \;\middle|\; f(x^*) < f(x) < f(x^*) + \eta \right\}$.

Do the generic objectives J above obey the KŁ property? YES

[Bolte, Daniilidis, Lewis, 2007]: If f is a subanalytic function with closed domain and continuous on its domain, then f has the KŁ property at each point of its domain.

Other examples: real-analytic and semi-algebraic functions

How to combine these facts in order to conclude about J?

Let f and g be two subanalytic functions. Then

(a) If f and g are lower-bounded, then $f + g$ is subanalytic.

(b) If g maps bounded sets on bounded sets or if $f^{-1}(\mathcal{X})$ is bounded for any bounded subset $\mathcal{X}$, then $f \circ g$ is a subanalytic function.

Global convergence using the Kurdyka-Łojasiewicz property

Assumption KŁ J has the KŁ property at a critical point $(x^*, y^*) \in \mathcal{L}(x^0, y^0)$

Proposition 4 (global convergence to a critical point) Let (x^k, y^k) be generated by ASAP under (H2)-(c-3-i) with $y \mapsto \nabla_x H(x, y)$ locally Lipschitz. Then,

(a) (x^k, y^k) is a Cauchy sequence that converges to (x^*, y^*)

(b) $\sum_{k=1}^{+\infty}(\|x^{k-1} - x^k\| + \|y^{k-1} - y^k\|) < +\infty$

Proof [Attouch, Bolte, Svaiter, 2013]

(Prop.1) $J(x^{k-1}, y^{k-1}) \geq J(x^k, y^k) + \rho\left(\|x^{k-1} - x^k\|^2 + \|y^{k-1} - y^k\|^2\right)$

(Prop.3) $\|(q_x^k, q_y^k)\| \leq \beta\|(x^{k-1} - x^k, y^{k-1} - y^k)\|$

(Assumption) $\exists k_j$ s.t. $(x^{k_j}, y^{k_j}) \to (x^*, y^*)$ and $J(x^{k_j}, y^{k_j}) \to J(x^*, y^*)$

Application: Fringe separation in interferometric images

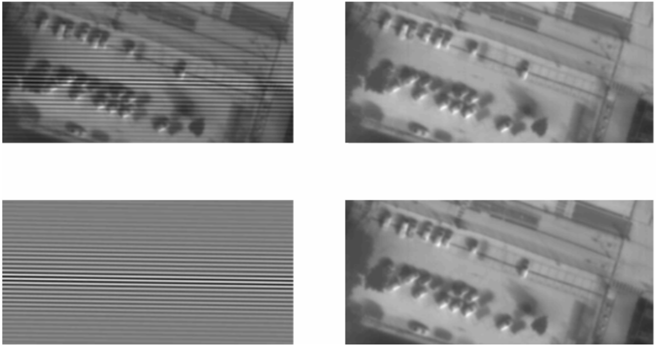
Model $w = x \circ y + \text{small noise}$

x: panchromatic image to recover
y: fringes to extract (known spectral support Ω and spatially structured)

$$J(x, y) = \mathrm{TV}^{\mathrm{v}}_{\alpha}(x) + \mathrm{TV}^{\mathrm{h}}_{\alpha}(y) + \|\mathbb{1}_{^c\Omega}\mathcal{F}(y)\|^2 + \frac{\mu}{2}\|x \circ y - w\|^2$$

TV^{h}: smoothed horizontal total variation
TV^{v}: smoothed vertical total variation
$\mathcal{F}$: Fourier transform

Sieleters airborne sequences $\approx$ thousands of 424×1000 images



Top left: data w, top right: panchromatic image x,
bottom left: fringes y, bottom right: reference

Applications: conventional stereo matching techniques, registration, hyperspectral infrared imagery

Fixed-pattern noise removing in interferometric images

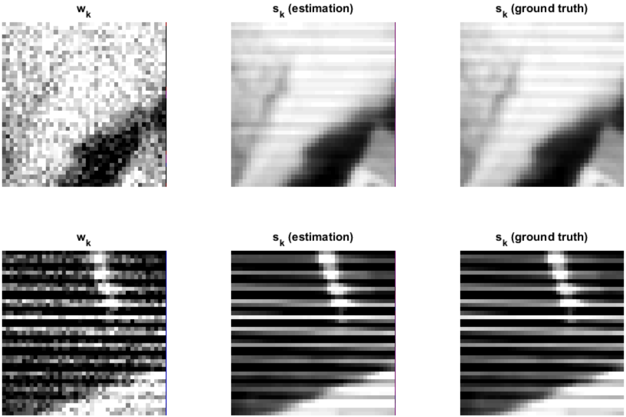
Nonuniform response of the infrared detector
(same input = different measures for different pixels)

Model $$w_k = (x_1 \circ u_k \circ y_k + x_2) + \text{small noise}$$

$x = (x_1, x_2)$: nonuniformity model parameters
$y = \{y_k\}$: fringes (known spectral support Ω and spatially structured)
u_k: known panchromatic images

$$J(x, y) = \sum_k \mathrm{TV}^{\mathrm{h}}_{\alpha}(y_k) + \sum_k \|\mathbb{1}_{^c\Omega} \mathcal{F}(y_k)\|^2 + \frac{\mu}{2} \sum_k \|x_1 \circ u_k \circ y_k + x_2 - w_k\|^2$$

TV^{h}: smoothed horizontal total variation
$\mathcal{F}$: Fourier transform



w_k
s_k (estimation)
s_k (ground truth)
w_k
s_k (estimation)
s_k (ground truth)

Application in image colorization

For any pixel p, C candidates $\{c_1(p), \cdots, c_C(p)\}$

$$J(x, y) = \mathrm{TV}_{\mathfrak{C}}(x) + \frac{\lambda}{2} \int_\Omega \sum_{i=1}^{C} y_i(p) \left\| x(p) - c_i(p) \right\|^2 \mathrm{d}p + \chi_{\mathcal{R}}(x) + \chi_{\Sigma}(y)$$

where $x = (U, V)$ are the chrominances to recover, in the range $\mathcal{R}$

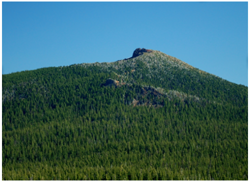
$y \in \Sigma$ iff $y_i(p) \geq 0$ and $\sum_{i=1}^{C} y_i(p) = 1$

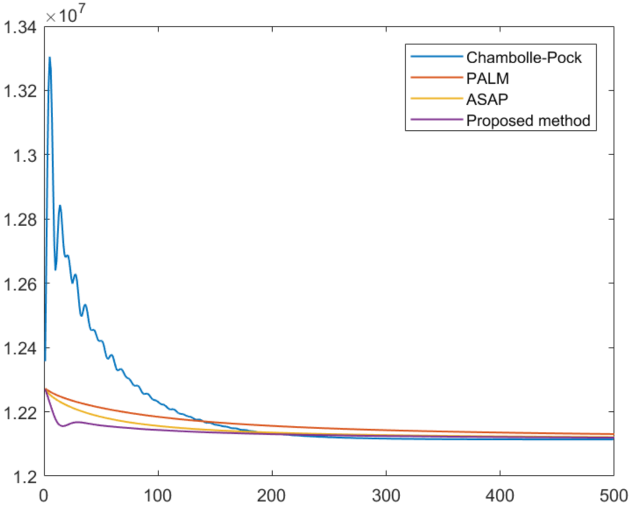
Y is the known luminance and

$$\mathrm{TV}_{\mathfrak{C}}(x) = \int_\Omega \sqrt{\gamma \|\nabla Y(p)\|^2 + \|\nabla U(p)\|^2 + \|\nabla V(p)\|^2} \, \mathrm{d}p$$

At convergence: $y_i(p) \in \{0, 1\}$ (vote process)
Colorization: $u(p) = c_i(p)$ iff $y_i(p) = 1$

Application in image colorization




×10^7
1.34
1.32
1.3
1.28
1.26
1.24
1.22
1.2
0
100
200
300
400
500
Chambolle-Pock
PALM
ASAP
Proposed method

Extensions

Proximal step: generalized proximity operators using Bregman distances (e.g. optimization on the simplex)

Acceleration: inertial (Nesterov) overrelaxation strategy (non-monotone scheme)

Some concluding notes

- The proposed ASAP is an alternative scheme to PALM for solving nonsmooth and nonconvex optimization problem
- Choice between ASAP and PALM depends on the structure and the regularity of the objective
- Biconvexity of the coupling term gives nice properties (large stepsizes)
- Promising applications on image processing
- Open questions: critical points vs. (local) minimum, initialization, theoretical convergence rate

Thank you for your attention!

Main references:

The ASAP algorithm: *Alternating structure-adapted proximal gradient descent for nonconvex block-regularised problems*
Mila Nikolova, P. T., *submitted*, 2017 (HAL-01677456)

Image colorization: *Inertial Alternating Generalized Forward-Backward Splitting for Image Colorization*
P.T., Fabien Pierre, Mila Nikolova, *submitted*, 2018 (HAL-01792432)

Fringe separation: *Fast and Accurate Multiplicative Decomposition for Fringe Removal in Interferometric Images*
Daniel Chen Soncco, Clara Barbanson, Mila Nikolova, Andrés Almansa, Yann Ferrec, IEEE Transactions on Computational Imaging, 2017

NU correction: *Correction par méthode variationnelle des non uniformités des détecteurs d'un interféromètre imageur*
P. T., Yann Ferrec, Laurent Rousset-Rouvière, colloque GRETSI, 2017