Mix Algorithms for Structured Convex Non-Differentiable Optimization

Cesare Molinari (Advisor: J. Peypouquet)

Doctorado en Matemáticas
Universidad Técnica Federico Santa María

Dubrovnik, 24 May 2017

Index

1. Introduction and previous works
   - General purpose
   - Optimality conditions and descent methods
   - Discretization: forward and backward
   - Forward-Backward
2. The constrained problem
   - 1) A primal algorithm: augmented energy
   - 2) A primal-dual algorithm: Lagrange multiplier
3. Applications
   - L^1/L^2 minimization
   - TV/L^2 minimization
4. Numerical example: sparse optimal control of PDE

We want to design and analyse iterative algorithms for the numerical resolution of optimization problems of the form

Structured optimization problem in two variables:

$$\min\{f(x) + g(y) : \; Ax + By = c\},$$

- X, Y, Z: Hilbert spaces; A, B lin. cont. op., $c \in Z$;
- $f : X \to \mathbb{R}$ convex and differentiable, $g : Y \to \bar{\mathbb{R}}$ convex and l.s.c.

Example: sparse optimal control of PDE

$x \in X$: control; $y \in Y$: state; $\mathcal{S} : X \to Y$: control-state map;

$$\min\{ \underbrace{\|x\|_{L^1}}_{\text{cost of control}} + \underbrace{\|y - y_d\|_{L^2}^2}_{\text{distance to target}} : \underbrace{Sx = y}_{\text{state equation}} \}$$

Index

General convex optimization problem

$$\tilde{x} \in Sol = \arg\min_{x \in X} f(x), \qquad (\mathcal{P})$$

- X Hilbert space;
- $f : X \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ proper, convex and l.s.c.

Theorem (Weierstrass-Hilbert-Tonelli, convex case)

In the previous setting, if f is *coercive* then $Sol \neq \emptyset$;
moreover, if f is *strictly-convex* then $Sol = \{\tilde{x}\}$.

Differentiable case

Hyp: $f \in C^1(X)$

Optimality condition (Fermat rule)

$$\tilde{x} \in Sol \quad \text{if and only if} \quad \nabla f(\tilde{x}) = 0$$

Dynamical system: gradient descent

$$\begin{cases} -\dot{x}(t) = \nabla f(x(t)) & \text{for } t \in (0, +\infty) \\ x(0) = x_0 \in X \end{cases} \qquad (\mathcal{DS})$$

Then $\frac{d}{dt} f(x(t)) = \langle \nabla f(x(t)), \dot{x}(t) \rangle = -\|\nabla f(x(t))\|^2 = -\|\dot{x}(t)\|^2$

Non-differentiable case: optimality condition

Theorem (Generalized Fermat rule)

$$\tilde{x} \in Sol \quad \text{if and only if} \quad 0 \in \partial f\left(\tilde{x}\right),$$

where the *sub-differential* in $\bar{x}$ is defined by

$$\partial f\left(\bar{x}\right) = \left\{x^* \in X : \; f\left(x\right) \geq f\left(\bar{x}\right) + \langle x^*, \; x - \bar{x}\rangle_X \; \text{ for every } x \in X\right\}$$

Non-differentiable case: $\mathcal{SDI}$

$$\begin{cases} -\dot{x}(t) \in \partial f(x(t)) & \text{for } t \in (0, +\infty) \\ x(0) = x_0 \in \overline{dom(f)} \end{cases} \qquad (\mathcal{SDI})$$

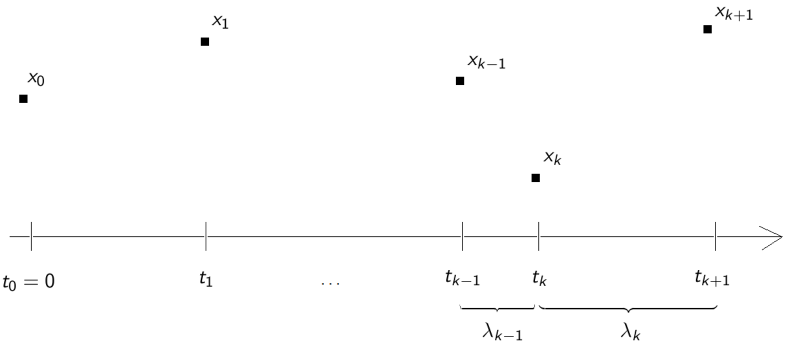
Theorem (Crandall and Pazy, 1969)

There exists a unique absolutely continuous function $x : [0, +\infty) \to X$ (∂f is a maximal monotone operator)

Idea: following the time-trajectory $x(t)$ until a *stationary point* (solution of the optimization problem)

Index

Discretization of $\mathcal{SDI}$



Remark

$$\lim_{k \to +\infty} t_k = +\infty \iff \lambda_k \notin \ell^1$$

Explicit discretization: the gradient method

Euler-forward

$$\dot{x}(t_k) \approx \frac{x_{k+1} - x_k}{\lambda_k}$$

- Hyp: f is differentiable in X $(\partial f(x) = \{\nabla f(x)\})$

Gradient method (explicit)

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k)$$

Implicit discretization: the proximal-point algorithm

Euler-backward

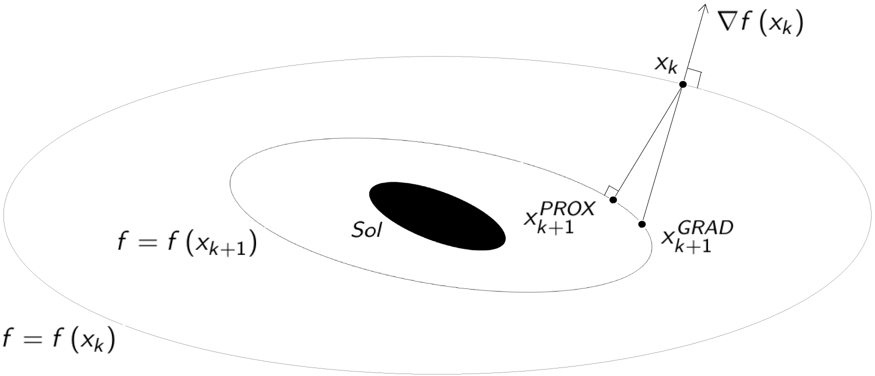
$$\dot{x}\left(t_{k+1}\right) \approx \frac{x_{k+1} - x_k}{\lambda_k}$$

Proximal point algorithm (Martinet, 1970)

$$-\frac{x_{k+1} - x_k}{\lambda_k} \in \partial f\left(x_{k+1}\right);$$

Equivalently, by Moreau-Rockafellar Theorem

$$x_{k+1} \in \arg\min_{x \in X} \left\{ f\left(x\right) + \frac{1}{2\lambda_k} \left\| x - x_k \right\|^2 \right\}$$



Remark (Grad VS Prox)

- Gradient method: $x_{k+1} - x_k \; // \; \nabla f(x_k)$ (explicit, but unstable);
- Proximal-point: $\langle x_{k+1} - x_k, \; x_k - x_{k-1} \rangle < 0$ (stable, but implicit)

Index

1. Introduction and previous works
   - General purpose
   - Optimality conditions and descent methods
   - Discretization: forward and backward
   - Forward-Backward
2. The constrained problem
   - 1) A primal algorithm: augmented energy
   - 2) A primal-dual algorithm: Lagrange multiplier
3. Applications
   - L^1/L^2 minimization
   - TV/L^2 minimization
4. Numerical example: sparse optimal control of PDE

Structured optimization problem in one variable:

$$x^* \in Sol = \arg\min_{x \in X} \{ \ F(x) = f(x) + g(x) \ \}, \qquad (\mathcal{P}_x^{mix})$$

- X: Hilbert space;
- $f : X \to \mathbb{R}$: convex and *differentiable* with L-Lipschitz gradient;
- $g : X \to \overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$: proper, convex and *l.s.c.*

Forward-backward splitting or ISTA (R. J. Bruck, 1977)

1) Gradient method on f: $x_{k+\frac{1}{2}} = x_k - \lambda_k \nabla f(x_k)$;

2) Proximal-point on g: $x_{k+1} = \arg\min \left\{ g(x) + \frac{1}{2\lambda_k} \|x - x_{k+\frac{1}{2}}\|^2 \right\}$

Example: Gradient-projected method

$$g(x) = \delta_C(x) = \begin{cases} 0 & \text{if: } x \in C \\ +\infty & \text{else,} \end{cases}$$

where $C \subset X$ non-empty, convex and closed; then

$$x_{k+1} = \Pi_C(x_k - \lambda_k \nabla f(x_k))$$

Mix algorithm: a different interpretation

$$x_{k+1} = \arg\min_{x \in X} \left\{ g(x) + \frac{1}{2\lambda_k} \| x - (x_k - \lambda_k \nabla f(x_k)) \|^2 \right\}$$

$$= \arg\min_{x \in X} \left\{ g(x) + \underbrace{f(x_k) + \langle \nabla f(x_k), \; x - x_k \rangle}_{\text{first-order Taylor expansion of } f \text{ at } x_k} + \underbrace{(1 \backslash 2\lambda_k) \| x - x_k \|^2}_{\text{distance penalization}} \right\}$$

Remark

Bigger is λ_k, larger is the distance allowed between x_{k+1} and x_k

Properties

Proposition

Hyp: $Sol \neq \emptyset$, $\lambda_n \notin \ell^1$ and $0 < \lambda_n \leq \lambda < \frac{2}{L}$;

i) $F(x_n) \searrow \inf_X \; F$

ii) if $x_{n_k} \rightharpoonup x$, then $x \in Sol$

iii) if $x \in Sol$, then $\lim_{n\to\infty} \|x_n - x\|$ exists finite

iv) $x_k \rightharpoonup \tilde{x} \in Sol$

Index

Linear-affine constrained two-variables problem

Structured optimization problem in two variables:

$$(x^*, y^*) \in Sol = \underset{(x,y)\in C}{\arg\min} \left\{ \, F(x,y) = f(x) + g(y) : \quad Ax + By = c \, \right\}, \qquad (\mathcal{P}_{xy}^{mix})$$

- X, Y, Z: Hilbert spaces;
- $f \in C^{1,1}(X; \mathbb{R}), g \in \Gamma_0\left(Y; \overline{\mathbb{R}}\right)$;
- A, B lin. cont. op., $c \in Z$

Index

1) Augmented energy (primal)

Well-posedness

$Sol \neq \emptyset$ and $(\tilde{x}, \tilde{y}) \in Sol$ if and only if $\exists \tilde{z} \in Z$ such that

$$\begin{cases} -A^* \tilde{z} = \nabla f(\tilde{x}) \\ -B^* \tilde{z} \in \partial g(\tilde{y}) \\ A\tilde{x} + B\tilde{y} = c \end{cases}$$

Idea: gradient method in x and proximal point in y applied to

Augmented energy

$$\mathcal{E}_\lambda(x, y) = f(x) + g(y) + \frac{\gamma}{2\lambda}\|Ax + By - c\|_Z^2$$

The algorithm $\mathcal{A}_{xy}^{aug}$

Process (in series): $(x_k, y_k) \longrightarrow (x_{k+1}, y_k) \longrightarrow (x_{k+1}, y_{k+1})$

1) Gradient step (in x variable):

$$x_{k+1} = x_k - \lambda_k \nabla f(x_k) - \gamma A^* (Ax_k + By_k - c);$$

2) Proximal-point step (in y variable):

$$y_{k+1} = \arg\min_{y \in Y} \left\{ g(y) + \frac{\gamma}{2\lambda_k} \|Ax_{k+1} + By - c\|_Z^2 + \frac{1}{2\lambda_k} \|y - y_k\|_Y^2 \right\}$$

Theorem (global weak convergence)

Hyp:

- Well-posedness
- $\lambda_k \in \ell^2$, decr. and $\sup_{k\in\mathbb{N}}\left(\frac{1}{\lambda_{k+1}} - \frac{1}{\lambda_k}\right) < +\infty$ (Ex: $\lambda_k = \frac{1}{k}$)
- $0 < \gamma < \frac{2}{\|A\|^2}$;

then $(x_k, y_k) \rightharpoonup (\tilde{x}, \tilde{y}) \in Sol$

Index

2) Lagrange multiplier (primal-dual)

Well-posedness (Saddle-point condition)

$Sol \neq \emptyset$ and $(\tilde{x}, \tilde{y}) \in Sol$ if and only if $\exists \tilde{p} \in Z$ such that

$$\mathcal{L}(\tilde{x}, \tilde{y}, p) \leq \mathcal{L}(\tilde{x}, \tilde{y}, \tilde{p}) \leq \mathcal{L}(x, y, \tilde{p}) \qquad \forall x \in X, y \in Y, p \in Z$$

Idea: Lagrange multiplier prediction, gradient method in x, proximal point in y and Lagrange multiplier correction on

Lagrangian functional

$$\mathcal{L}(x, y, p) = f(x) + g(y) + \langle p, Ax + By - c\rangle$$

The Algorithm $\mathcal{A}_{xy}^{lag}$

Process (series, parallel, series):

$$(x_k, y_k, p_k) \to (x_k, y_k, p_{k+1}) \to (x_{k+1}, y_{k+1}, p_{k+1}) \to (x_{k+1}, y_{k+1}, z_{k+1})$$

1) Max-gradient in p (Lagrange multiplier *Prediction*):

$$\begin{aligned} p_{k+1} &= z_k + \lambda_k \nabla_p \, \mathcal{L}(x_k, y_k, z_k) \\ &= z_k + \lambda_k (Ax_k + By_k - c); \end{aligned}$$

2) Min-gradient in x:

$$\begin{aligned} x_{k+1} &= x_k - \lambda_k \nabla_x \, \mathcal{L}(x_k, y_k, p_{k+1}) \\ &= x_k - \lambda_k [\nabla f(x_k) + A^* p_{k+1}]; \end{aligned}$$

3) Proximal-point in y:

$$y_{k+1} = \arg\min_{y \in Y} \left\{ \mathcal{L}\left(x_k, y, p_{k+1}\right) + \frac{1}{2\lambda_k} \|y - y_k\|_Y^2 \right\}$$
$$= \arg\min_{y \in Y} \left\{ g\left(y\right) + \frac{1}{2\lambda_k} \|y - \tilde{y}_k\|_Y^2 \right\}, \quad \text{for } \tilde{y}_k = y_k - \lambda_k B^* p_{k+1};$$

4) Max-gradient in p (Lagrange multiplier *Correction*):

$$z_{k+1} = z_k + \lambda_k \left(Ax_{k+1} + By_{k+1} - c\right)$$

The convergence result

Theorem (weak global convergence)

Hyp: Well-posedness and $\exists \lambda_{min} > 0$ such that

$$0 < \lambda_{min} \leq \lambda_k \leq \lambda_{max} = \text{func}\,(L, \|A\|, \|B\|)\,;$$

then $(x_k, y_k, p_k) \rightharpoonup (\tilde{x}, \tilde{y}, \tilde{p}) \in Sol$

Index

Unconstrained LASSO (R. Tibshirani, 1996) for statistical regression:

$$\boxed{\min_{u\in L^2}\left\{\frac{1}{2}\|Au-\bar{u}\|_{L^2}^2+\mu\|u\|_{L^1}\right\}}$$

<u>Difficulties</u>:

- non-differentiability of $\|\cdot\|_{L^1}$
- coordinates "mixed" by operator A

<u>BUT</u> *separable structure* of the objective: $F=f+g$ with $f\in C^{1,1}$

Forward-backward $\left(\mathcal{A}_x^{mix}\right)$

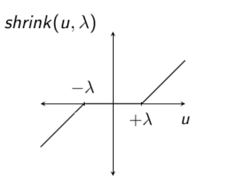
1) Gradient step: $u_{k+\frac{1}{2}} = u_k - \lambda_k A^* \left(Au_k - \bar{u}\right);$

2) Proximal-point: $u_{k+1} \in \arg\min_{L^2} \left\{ \|u\|_{L^1} + \frac{1}{2\mu\lambda_k} \|u - u_{k+\frac{1}{2}}\|_{L^2}^2 \right\}$

Soft-shrinkage

Proximal step has explicit solution: $u_{k+1} = shrink(u_{k+\frac{1}{2}}, \mu\lambda_k)$ for

$$shrink\,(u, \lambda) = \begin{cases} u + \lambda & (\text{if } u < -\lambda) \\ 0 & (\text{if } -\lambda \leq u \leq \lambda) \\ u - \lambda & (\text{if } u > \lambda) \end{cases}$$



Index

ROF model (Rudin Osher and Fatemi, 1992) for images reconstruction:

$$\boxed{\min_{u\in H^1}\left\{\frac{1}{2}\|Au-\bar{u}\|_{L^2}^2+\mu\|\nabla u\|_{L^1}\right\}}$$

Remark

Not only the operator A but also the gradient "mixes the coordinates"; in particular, using ISTA the proximal-point has not explicit solution:

$$u_{k+1}\in\arg\min_{u\in H^1}\left\{\|\nabla u\|_{L^1}+\frac{1}{2\lambda_k}\|u-u_{k+\frac{1}{2}}\|_{L^2}^2\right\}$$

The equivalent constrained problem

Idea: introducing an *auxiliary variable*, we can obtain an *equivalent problem* of the form $\mathcal{P}_{xy}^{mix}$:

$$\min_{(u,\mathbf{q})\in H^1\times L^2}\left\{\|\mathbf{q}\|_{L^1}+\frac{1}{2\mu}\|Au-\bar{u}\|_{L^2}^2\,:\;\mathbf{q}=\nabla u\right\}$$

TV minimization through $\mathcal{A}_{xy}^{aug}$

1) Gradient step:

$$u_{k+1} = u_k - \frac{\lambda_k}{2\mu} \nabla_{H^1} \left(\|Au - \bar{u}\|_{L^2}^2 \right)(u_k) - \gamma \nabla^* \left(\nabla u_k - \mathbf{q}_k \right);$$

2) Proximal-point:

$$\mathbf{q}_{k+1} = \arg\min_{\mathbf{q} \in L^2} \left\{ \|\mathbf{q}\|_{L^1} + \frac{\gamma}{2\lambda_k} \|\nabla u_{k+1} - \mathbf{q}\|_{L^2}^2 + \frac{1}{2\lambda_k} \|\mathbf{q} - \mathbf{q}_k\|_{L^2}^2 \right\}$$

$$= \arg\min_{\mathbf{q} \in L^2} \left\{ \|\mathbf{q}\|_{L^1} + \frac{1}{2\delta_k} \|\mathbf{q} - \tilde{\mathbf{q}}_k\|_{L^2}^2 \right\} = shrink\left(\tilde{\mathbf{q}}_k, \delta_k \right),$$

for $\delta_k = \frac{\lambda_k}{1+\gamma}$ and $\tilde{\mathbf{q}}_k = \frac{\mathbf{q}_k + \gamma \nabla u_{k+1}}{1+\gamma}$

Index

Optimal control of elliptic PDE

The continuous problem

$$\min_{u,y\in L^2(\Omega)} \left\{ \frac{1}{2}\|y - y_d\|_{L^2}^2 + \alpha\|u\|_{L^1} \right\},$$

$$\begin{cases} -\nu\Delta y + ky = f + u & (\Omega) \\ y = 0 & (\partial\Omega) \end{cases}$$

Discretize (through Finite Elements), then optimize

$$\min_{\mathbf{u},\mathbf{y}\in\mathbb{R}^N} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{y}_d\|_2^2 + \alpha\|\mathbf{u}\|_1 \right\},$$

$$B\mathbf{y} - M\mathbf{u} = \mathbf{f}$$

Remark: ℓ^1-norm promotes sparsity, BUT it is not-differentiable!

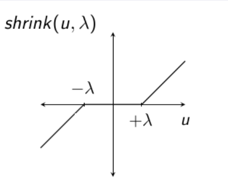
Mix algorithm for optimal control

- Remark 1: during the algorithm execution, it is not required the inversion of matrix B (i.e.: to solve numerically the PDE);
- Remark 2: the "implicit" proximal-point step has explicit solution:

$$u_{k+1} = shrink(u_k + \lambda_k M\tilde{\mu}_k, \alpha\lambda_k), \quad \text{where}$$

Soft-shrinkage operator

$$shrink\,(u,\lambda) = \begin{cases} u+\lambda & (\text{if } u < -\lambda) \\ 0 & (\text{if } -\lambda \le u \le \lambda) \\ u-\lambda & (\text{if } u > \lambda) \end{cases}$$



Numerical results: final solution

$\Omega = (0, 1)$; $\nu = 1, k = 0.001$; $\alpha = 0.002$; $N_{nodes} = 200$; $N_{it} = 7e6$
$u_0 = 0, y_0 = y_{omog} = \sin(2\pi x)$; $\lambda = 0.04$; $\gamma = 0.001$

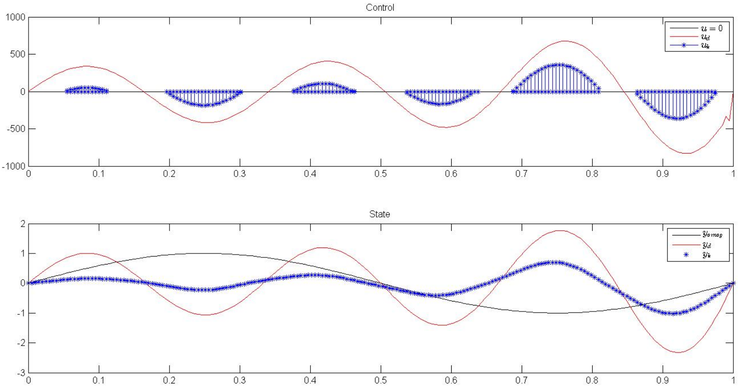


Figure: $u = 0, y_{omog}$ (black); u_d, y_d (red); $\bar{u}, \bar{y}$ (blue)

Numerical results: energy and constraint

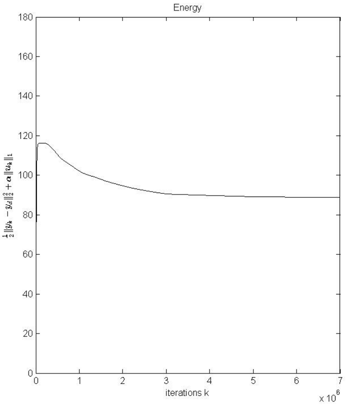


Figure: Energy $\frac{1}{2}\|y_k - y_d\|_2^2 + \alpha\|u_k\|_1$ and gap $\|By_k - Mu_k - f\|_2$

Essential bibliography

- H. Brézis, *Opérateurs Maximaux Monotones et Semi-Groupes de Contractions dans les Espaces de Hilbert*, North Holland (1973)
- B. Martinet, *Régularisation d'inéquations variationnelles par approximations successives*, Revue française d'informatique et recherche opérationnelle (1970)
- G. Chen, M. Teboulle, *A proximal-based decomposition method for convex minimization problems*, Mathematical Programming (1994)
- A. Beck, M. Teboulle, A Fast Iterative Shrinkage-Thresholding Algorithm for linear inverse problem, SIAM J. Imaging Sciences (2009)
- L. I. Rudin, S. Osher, E. Fatemi, *Nonlinear Total Variation Based Noise Removal Algorithms*, Physica D (1992)
- R. Tibshirani, *Regression shrinkage and selection via the LASSO*, Journal of the Royal Statistical Society (1996)