A primal-dual framework for mixtures of regularizers

Baran Gözcü
baran.goezcue@epfl.ch

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)
Switzerland

EPFL
[March 25, 2015]

Joint work with

Luca Baldassarre, Quoc Tran Dinh, Cosimo Aprile and Volkan Cevher @ LIONS

lions@epfl
MARIE CURIE ACTIONS
FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION
erc
EPFL

Outline

Overview of compressive imaging

System model

$$\mathbf{y} = M\mathbf{x} + \boldsymbol{\omega} \qquad (1)$$

- M is the measurement matrix (Fourier, Gaussian etc.)
- $\mathbf{x}$ is the image that is in vectorized form
- $\boldsymbol{\omega}$ is the image that is in vectorized form
- $\mathbf{b}$ is the measurement vector

Solution

Then we solve

$$\begin{array}{ll} \min_{\mathbf{x}} & \|W\mathbf{x}\|_1 \\ \text{subject to} & M\mathbf{x} = \mathbf{y} \end{array} \qquad (2)$$

where W is the sparsity basis.

Why a mixture model ?

- Signals can posseses various structures at the same time.
- For example the reconstruction with TV norm $\|\mathbf{x}\|_{TV} = \sum_{i,j} \|(\nabla(\mathbf{x}))_{i,j}\|_2$

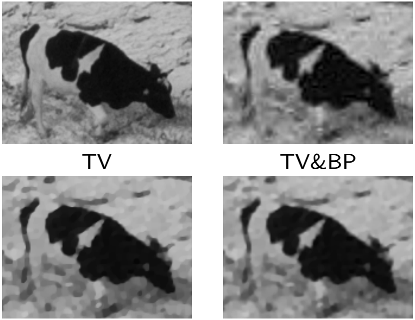
$$\begin{array}{ll} \min_{\mathbf{x}} & \|\mathbf{x}\|_{TV} \\ \text{subject to} & M\mathbf{x} = \mathbf{y} \end{array} \tag{3}$$

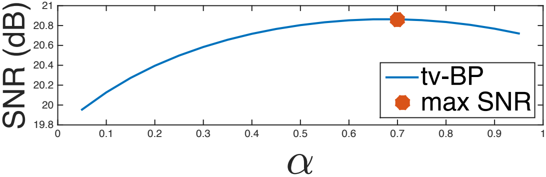
  will introduce flat regions.
- What happens if we solve with a mixture of regularizers ?

$$\begin{array}{ll} \min_{\mathbf{x}} & \alpha\|\mathbf{x}\|_{TV} + (1-\alpha)\|\mathbf{x}\|_1 \\ \text{subject to} & M\mathbf{x} = \mathbf{y} \end{array} \tag{4}$$

Illustration: Mixture model performs better

Original (2048x2048,%15 of DCT)





General Problem

$$
\begin{array}{ll}
\min_{\mathbf{x}} & f(\mathbf{x}) := \sum_{i=0}^{p} f_i(\mathbf{x}) \\
\text{subject to} & A\mathbf{x} = \mathbf{b}
\end{array}
\tag{5}
$$

How to solve with

- computational efficiency
- guarantee on objective function
- guarantee on feasibility

Outline

Mixture of regularizers

Constrained convex minimization: A primal-dual framework

Application

Conclusion

Swiss army knife of convex formulations

Our primal problem prototype: A simple mathematical formulation[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\}, \quad (6)$$

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (6) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

[1] We can simply replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{C}$ for a convex cone $\mathcal{C}$ without any fundamental change.

Swiss army knife of convex formulations

Our **primal problem** prototype: A simple mathematical formulation[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \Big\}, \tag{6}$$

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (6) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

Example to keep in mind in the sequel

$$\mathbf{x}^\star := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ \|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}, \|\mathbf{x}\|_\infty \leq 1 \Big\}$$

[1] We can simply replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{C}$ for a convex cone $\mathcal{C}$ without any fundamental change.

Swiss army knife of convex formulations

Our **primal problem** prototype: A simple mathematical formulation[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \Big\}, \tag{6}$$

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (6) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

Example to keep in mind in the sequel

$$\mathbf{x}^\star := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ \|\mathbf{x}\|_1 : \mathbf{A}\mathbf{x} = \mathbf{b}, \|\mathbf{x}\|_\infty \leq 1 \Big\}$$

Broader context for (6):

- Standard convex optimization formulations: *linear programming, convex quadratic programming, second order cone programming, semidefinite programming and interior point algorithms*.
- Reformulations of existing unconstrained problems via **convex splitting**: *composite convex minimization, consensus optimization, ...*

[1] We can simply replace $\mathbf{A}\mathbf{x} = \mathbf{b}$ with $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{C}$ for a convex cone $\mathcal{C}$ without any fundamental change.

Numerical ϵ-accuracy

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (6) is **impracticable** unless problem has a closed form solution, which is extremely limited in reality.
- Numerical optimization algorithms result in $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to a given accuracy ϵ in **some sense**.

- In the sequel, by ϵ-accurate solutions $\mathbf{x}^\star_\epsilon$ of (6), we mean the following

Definition (ϵ-accurate solutions)

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (6) if

$$\begin{cases} |f(\mathbf{x}^\star_\epsilon) - f^\star| \leq \epsilon & \text{(objective residual)}, \\ \|\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}\| \leq \epsilon & \text{(feasibility gap)}, \\ \mathbf{x}^\star_\epsilon \in \mathcal{X} & \text{(exact simple set feasibility)}. \end{cases}$$

- Indeed, ϵ can be different for the objective, feasibility gap, or the iterate residual.

The optimal solution set

Before we talk about algorithms, we must first characterize what we are looking for!

Optimality condition

The **optimality condition** of $\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ can be written as

$$\begin{cases} 0 & \in \mathbf{A}^T\lambda^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases} \tag{7}$$

(Subdifferential) $\partial f(\mathbf{x}) := \{\mathbf{v} \in \mathbb{R}^p \ : \ f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{v}^T(\mathbf{y}-\mathbf{x}), \ \forall \mathbf{y} \in \mathbb{R}^p\}$.

- This is the well-known KKT (Karush-Kuhn-Tucker) condition.
- Any point $(\mathbf{x}^\star, \lambda^\star)$ satisfying (7) is called a KKT point.
- $\mathbf{x}^\star$ is called a stationary point and $\lambda^\star$ is the corresponding multipliers.

Lagrange function and the minimax formulation

We can naturally interpret the optimality condition via a minimax formulation

$$\max_{\lambda} \min_{\mathbf{x}\in \mathrm{dom}(f)} \mathcal{L}(\mathbf{x}, \lambda),$$

where $\lambda \in \mathbb{R}^n$ is the vector of Lagrange multipliers or dual variables w.r.t. $\mathbf{A}\mathbf{x} = \mathbf{b}$ associated with the Lagrange function:

$$\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x} - \mathbf{b})$$

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (6), we therefore seek the solutions

$$(\mathbf{x}^\star, \lambda^\star) \in \arg\max_{\lambda} \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x}, \lambda),$$

which we can naively brake down into two—in general **nonsmooth**—problems:

$$\begin{array}{lll} \textbf{Lagrangian subproblem:} & \mathbf{x}^*(\lambda) & \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\} \\ \textbf{Dual problem:} & \lambda^\star & \in \arg\max_{\lambda} \{d(\lambda) := \mathcal{L}(\mathbf{x}^*(\lambda), \lambda)\} \end{array}$$

- The function $d(\lambda)$ is called the dual function.
- The optimal dual objective value is $d^\star = d(\lambda^\star)$.

The dual function $d(\lambda)$ is concave. Hence, we can attempt the following strategy:

1. Find the optimal solution $\lambda^\star$ of the "convex" dual problem.
2. Obtain the optimal primal solution $\mathbf{x}^\star = \mathbf{x}^*(\lambda^\star)$ via the convex primal problem.

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (6), we therefore seek the solutions

$$(\mathbf{x}^\star, \lambda^\star) \in \arg\max_{\lambda} \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x}, \lambda),$$

which we can naively brake down into two—in general **nonsmooth**—problems:

Lagrangian subproblem: $\mathbf{x}^*(\lambda) \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\}$

Dual problem: $\lambda^\star \in \arg\max_{\lambda} \{d(\lambda) := \mathcal{L}(\mathbf{x}^*(\lambda), \lambda)\}$

- The function $d(\lambda)$ is called the dual function.
- The optimal dual objective value is $d^\star = d(\lambda^\star)$.

The dual function $d(\lambda)$ is concave. Hence, we can attempt the following strategy:

1. Find the optimal solution $\lambda^\star$ of the "convex" dual problem.
2. Obtain the optimal primal solution $\mathbf{x}^\star = \mathbf{x}^*(\lambda^\star)$ via the convex primal problem.

Challenges for the plausible strategy above

1. Establishing its correctness
2. Computational efficiency of finding an $\bar{\epsilon}$-approximate optimal dual solution $\lambda^\star_{\bar{\epsilon}}$
3. Mapping $\lambda^\star_{\bar{\epsilon}} \to \mathbf{x}^\star_{\epsilon}$ (i.e., $\bar{\epsilon}(\epsilon)$), where ϵ is for the original constrained problem (6)

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (6), we therefore seek the solutions

$$(\mathbf{x}^\star, \lambda^\star) \in \arg\max_{\lambda} \min_{\mathbf{x}\in\mathcal{X}} \mathcal{L}(\mathbf{x}, \lambda),$$

which we can naively brake down into two—in general **nonsmooth**—problems:

Lagrangian subproblem: $\mathbf{x}^*(\lambda) \in \arg\min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\}$

Dual problem: $\lambda^\star \in \arg\max_{\lambda} \{d(\lambda) := \mathcal{L}(\mathbf{x}^*(\lambda), \lambda)\}$

- The function $d(\lambda)$ is called the dual function.
- The optimal dual objective value is $d^\star = d(\lambda^\star)$.

The dual function $d(\lambda)$ is concave. Hence, we can attempt the following strategy:

1. Find the optimal solution $\lambda^\star$ of the "convex" dual problem.
2. Obtain the optimal primal solution $\mathbf{x}^\star = \mathbf{x}^*(\lambda^\star)$ via the convex primal problem.

Challenges for the plausible strategy above

1. Establishing its correctness: Assume $f^\star > -\infty$ and Slater's condition for $f^\star = d^\star$
2. Computational efficiency of finding an $\bar{\epsilon}$-approximate optimal dual solution $\lambda^\star_{\bar{\epsilon}}$
3. Mapping $\lambda^\star_{\bar{\epsilon}} \to \mathbf{x}^\star_{\epsilon}$ (i.e., $\bar{\epsilon}(\epsilon)$), where ϵ is for the original constrained problem (6)

Nesterov's smoothing idea: From $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ to $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$

When can the dual function have Lipschitz gradient?

When $f(\mathbf{x})$ is γ-strongly convex, the dual function $d(\lambda)$ is $\frac{\|\mathbf{A}\|^2}{\gamma}$-Lipschitz gradient.

(Strong convexity) $f(\mathbf{x})$ is γ-strongly convex iff $f(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x}\|_2^2$ is convex.

$$d(\lambda) = \min_{\mathbf{x}:\mathbf{x}\in\mathcal{X}} \underbrace{f(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x}\|_2^2}_{\text{convex \& possibly nonsmooth}} + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \underbrace{\frac{\gamma}{2}\|\mathbf{x}\|_2^2}_{\text{leads to } d\in\mathcal{F}_L}$$

AGM automatically obtains $d^\star - d(\mathbf{x}^k) \leq \bar{\epsilon}$ with $k = \mathcal{O}\left(\frac{1}{\sqrt{\bar{\epsilon}}}\right)$

Nesterov's smoothing idea: From $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{1}{\epsilon}\right)$

When can the dual function have Lipschitz gradient?

When $f(\mathbf{x})$ is γ-strongly convex, the dual function $d(\lambda)$ is $\frac{\|\mathbf{A}\|^2}{\gamma}$-Lipschitz gradient.

(Strong convexity) $f(\mathbf{x})$ is γ-strongly convex iff $f(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x}\|_2^2$ is convex.

$$d(\lambda) = \min_{\mathbf{x}:\mathbf{x}\in\mathcal{X}} \underbrace{f(\mathbf{x}) - \frac{\gamma}{2}\|\mathbf{x}\|_2^2}_{\text{convex \& possibly nonsmooth}} + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \underbrace{\frac{\gamma}{2}\|\mathbf{x}\|_2^2}_{\text{leads to } d\in\mathcal{F}_L}$$

Nesterov's smoother [3]

We add a strongly convex term to Lagrange subproblem so that the dual is smooth!

$$d_\gamma(\lambda) = \min_{\mathbf{x}:\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_c\|_2^2, \text{with a center point } \mathbf{x}_c \in \mathcal{X}$$

$\nabla d_\gamma(\lambda) = \mathbf{A}\mathbf{x}^*_\gamma(\lambda) - \mathbf{b}$ ($\mathbf{x}^*_\gamma(\lambda)$: the γ-Lagrangian subproblem solution)

1. $d_\gamma(\lambda) - \gamma\mathcal{D}_\mathcal{X} \leq d(\lambda) \leq d_\gamma(\lambda)$, where $\mathcal{D}_\mathcal{X} = \max_{\mathbf{x}\in\mathcal{X}} \frac{1}{2}\|\mathbf{x} - \mathbf{x}_c\|_2^2$.
2. $\mathbf{x}^k$ of AGM on $d_\gamma(\lambda)$ has $d^\star - d(\mathbf{x}^k) \leq \gamma\mathcal{D}_\mathcal{X} + d^\star_\gamma - d_\gamma(\mathbf{x}^k) \leq \gamma\mathcal{D}_\mathcal{X} + \frac{2\|\mathbf{A}\|^2 R^2}{\gamma(k+2)^2}$.
3. We minimize the upperbound wrt γ and obtain $d^\star - d(\mathbf{x}^k) \leq \bar{\epsilon}$ with $k = \mathcal{O}\left(\frac{1}{\epsilon}\right)$.

Computational efficiency: The key role of the prox-operator

Smoothed dual: $d_\gamma(\lambda) = \min_{\mathbf{x}:\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \lambda, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{\gamma}{2}\|\mathbf{x} - \mathbf{x}_c\|_2^2$

$$\mathbf{x}^*(\lambda) = \mathrm{prox}_{f/\gamma}\left(\mathbf{x}_c - \frac{1}{\gamma}\mathbf{A}^T\lambda\right)$$

Definition (Prox-operator)

$$\mathrm{prox}_g(\mathbf{x}) := \arg\min_{\mathbf{z}\in\mathbb{R}^p}\{g(\mathbf{z}) + (1/2)\|\mathbf{z} - \mathbf{x}\|^2\}.$$

Key properties:

- distributes when the primal problem has decomposable structure:

$$f(\mathbf{x}) := \sum_{i=1}^{m} f_i(\mathbf{x}_i), \quad \text{and} \quad \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_m.$$

  where $m \geq 1$ is the number of components.
- often efficient & has closed form expression. For instance, if $g(\mathbf{z}) = \|\mathbf{z}\|_1$, then the prox-operator performs coordinate-wise soft-thresholding by 1.

Going from the dual $\bar{\epsilon}$ to the primal ϵ–I

Optimality condition (revisted)

Two equivalent ways of viewing the optimality condition of the primal problem (6)

mixed variational inequality (MVIP) **inclusion**

$$\boxed{f(\mathbf{x}) - f(\mathbf{x}^\star) + M(\mathbf{z}^\star)^T(\mathbf{z} - \mathbf{z}^\star) \geq 0, \quad \forall \mathbf{z} \in \mathcal{X} \times \mathbb{R}^n} = \begin{cases} 0 & \in \mathbf{A}^T\lambda^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases}$$

where $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}$ and $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (6).

Going from the dual $\bar{\epsilon}$ to the primal ϵ–I

Optimality condition (revisted)

Two equivalent ways of viewing the optimality condition of the primal problem (6)

mixed variational inequality (MVIP) **inclusion**

$$\boxed{f(\mathbf{x}) - f(\mathbf{x}^\star) + M(\mathbf{z}^\star)^T(\mathbf{z} - \mathbf{z}^\star) \geq 0, \quad \forall \mathbf{z} \in \mathcal{X} \times \mathbb{R}^n} = \begin{cases} 0 & \in \mathbf{A}^T\lambda^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases}$$

where $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}$ and $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (6).

Measuring progress via the gap function

Unfortunately, measuring progress with the inclusion formulation is hard. However, associated with MVIP, we can define a **gap function** to measure our progress

$$G(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{X} \times \mathbb{R}^n} \left\{ f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) \right\}. \tag{8}$$

Key observations:

- $G(\mathbf{z}) = \underbrace{\max_{\hat{\lambda} \in \mathbb{R}^n} f(\mathbf{x}) + \langle \hat{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle}_{= f(\mathbf{x}) \text{ if } \mathbf{A}\mathbf{x} = \mathbf{b}, \infty \text{ o/w}} - \underbrace{\min_{\hat{\mathbf{x}} \in \mathcal{X}} f(\hat{\mathbf{x}}) + \langle \lambda, \mathbf{A}\hat{\mathbf{x}} - \mathbf{b} \rangle}_{= d(\lambda)} \geq 0, \forall \mathbf{z} \in \mathcal{X} \times \mathbb{R}^n$
- $G(\mathbf{z}^\star) = 0$ iff $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (6).
- Primal accuracy ϵ and the dual accuracy $\bar{\epsilon}$ can be related via the gap function.

Going from the dual $\bar{\epsilon}$ to the primal ϵ–II

A smoothed gap function measuring the excessive primal-dual gap

We define a smoothed version of the gap function $G_{\gamma\beta}(\mathbf{z}) =$

$$\underbrace{\max_{\hat{\lambda}\in\mathbb{R}^n} f(\mathbf{x}) + \langle \hat{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - \frac{\beta}{2}\|\hat{\lambda} - \hat{\lambda}_c\|_2^2}_{= f_\beta(\mathbf{x}) = f(\mathbf{x}) + \langle \hat{\lambda}_c, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2} - \underbrace{\min_{\hat{\mathbf{x}}\in\mathcal{X}} f(\hat{\mathbf{x}}) + \langle \lambda, \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\rangle + \frac{\gamma}{2}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_c\|_2^2}_{= d_\gamma(\lambda)}$$

where $(\hat{\mathbf{x}}_c, \hat{\lambda}_c) \in \mathcal{X} \times \mathbb{R}^n$ are primal-dual center points. In the sequel, they are 0.

- The primal accuracy ϵ is related to our primal model estimate $f_\beta(\mathbf{x})$
- The dual accuracy $\bar{\epsilon}$ is related to our smoothed dual function $d_\gamma(\lambda)$
- We must relate $G_{\gamma\beta}(\mathbf{z})$ to $G(\mathbf{z})$ so that we can tie ϵ to $\bar{\epsilon}$

Going from the dual $\bar{\epsilon}$ to the primal ϵ–II

A smoothed gap function measuring the excessive primal-dual gap

We define a smoothed version of the gap function $G_{\gamma\beta}(\mathbf{z}) =$

$$\underbrace{\max_{\hat{\lambda}\in\mathbb{R}^n} f(\mathbf{x}) + \langle\hat{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle - \frac{\beta}{2}\|\hat{\lambda} - \hat{\lambda}_c\|_2^2}_{= f_\beta(\mathbf{x}) = f(\mathbf{x}) + \langle\hat{\lambda}_c, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle + \frac{1}{2\beta}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2} - \underbrace{\min_{\hat{\mathbf{x}}\in\mathcal{X}} f(\hat{\mathbf{x}}) + \langle\lambda, \mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\rangle + \frac{\gamma}{2}\|\hat{\mathbf{x}} - \hat{\mathbf{x}}_c\|_2^2}_{= d_\gamma(\lambda)}$$

where $(\hat{\mathbf{x}}_c, \hat{\lambda}_c) \in \mathcal{X} \times \mathbb{R}^n$ are primal-dual center points. In the sequel, they are 0.

- The primal accuracy ϵ is related to our primal model estimate $f_\beta(\mathbf{x})$
- The dual accuracy $\bar{\epsilon}$ is related to our smoothed dual function $d_\gamma(\lambda)$
- We must relate $G_{\gamma\beta}(\mathbf{z})$ to $G(\mathbf{z})$ so that we can tie ϵ to $\bar{\epsilon}$

Our algorithm via MEG: model-based excessive gap (cf., [4])

Let $G_k(\cdot) := G_{\gamma_k\beta_k}(\cdot)$. We generate a **sequence** $\{\bar{\mathbf{z}}^k, \gamma_k, \beta_k\}_{k\geq 0}$ such that

$$\boxed{G_{k+1}(\bar{\mathbf{z}}^{k+1}) \leq (1 - \tau_k) G_k(\bar{\mathbf{z}}^k) + \psi_k} \qquad \text{(MEG)}$$

for $\psi_k \to 0$, rate $\tau_k \in (0,1)$ $(\sum_k \tau_k = \infty)$, $\gamma_k\beta_{k+1} < \gamma_k\beta_k$ so that $G_{\gamma_k\beta_k}(\cdot) \to G(\cdot)$.

- **Consequence**: $\boxed{G(\bar{\mathbf{z}}^k) \to 0^+ \quad \Rightarrow \quad \bar{\mathbf{z}}^k \to \mathbf{z}^\star = (\mathbf{x}^\star, \lambda^\star)}$ (primal-dual solution).

Going from the dual $\bar{\epsilon}$ to the primal ϵ–III

An uncertainty relation via MEG

The product of the primal and dual convergence rates is lowerbounded by MEG:

$$\boxed{\gamma_k \beta_k \geq \tau_k^2 \|\mathbf{A}\|^2}$$

Note that $\tau_k^2 = \Omega\left(\frac{1}{k^2}\right)$ due to Nesterov's lowerbound.

- The rate of γ_k controls the primal residual: $|f(\mathbf{x}^k) - f^\star| \leq \mathcal{O}(\gamma_k)$
- The rate of β_k controls the feasibility: $\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|_2 \leq \mathcal{O}(\beta_k + \tau_k) = \mathcal{O}(\beta_k)$
- They cannot be simultaneously $\mathcal{O}\left(\frac{1}{k^2}\right)$!

Convergence guarantee

Theorem [4, 5]

1. When f is strongly convex with $\mu > 0$, we can take $\gamma_k = \mu$ and $\beta_k = \mathcal{O}\left(\frac{1}{k^2}\right)$:

$$\begin{cases} -D_{\Lambda^\star}\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| \leq & f(\mathbf{x}^k) - f^\star & \leq 0 \\ & \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| & \leq \frac{4\|\mathbf{A}\|^2}{(k+2)^2\mu} D_{\Lambda^\star} \\ & \|\mathbf{x}^k - \mathbf{x}^\star\| & \leq \frac{4\|\mathbf{A}\|}{(k+2)\mu} D_{\Lambda^\star} \end{cases}$$

2. When f is non-smooth, the best we can do is $\gamma_k = \mathcal{O}\left(\frac{1}{k}\right)$ and $\beta_k = \mathcal{O}\left(\frac{1}{k}\right)$:

$$\begin{cases} -D_{\Lambda^\star}\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| \leq & f(\mathbf{x}^k) - f^\star & \leq \frac{2\sqrt{2}\|\mathbf{A}\| D_{\mathcal{X}}}{k+1}, \\ & \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| & \leq \frac{2\sqrt{2}\|\mathbf{A}\|(D_{\Lambda^\star} + \sqrt{D_{\mathcal{X}}})}{k+1}. \end{cases}$$

Outline

An application: Magnetic Resonance Imaging (MRI)

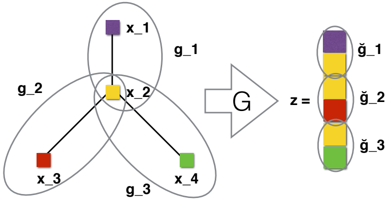
Mixture Model:

$$\min_{\mathbf{x}} \frac{1}{2}\|M\mathbf{x}-\mathbf{y}\|_2^2 + \alpha\|\mathbf{x}\|_{\mathsf{TV}} + \mu\|W\mathbf{x}\|_1 + \beta\|W\mathbf{x}\|_{\text{tree}} \tag{9}$$

$$\|\mathbf{x}\|_{\text{tree}} := \sum_{i=1}^{s} \|\mathbf{x}_{g_i}\|_2 \tag{10}$$

With $\mathbf{z} = G\mathbf{x}$ we can define the tree norm with non-overlapping groups $\tilde{g}_i$

$$\|\mathbf{x}\|_{\text{tree}} = \sum_{i=1}^{s} \|\left(G\mathbf{x}\right)_{\tilde{g}_i}\|_2 \tag{11}$$



***Wavelet Tree Sparsity Algorithm* (WaTMRI) [1]**

Mixture Model

$$\min_{\mathbf{x},\mathbf{z}} \frac{1}{2}\|M\mathbf{x}-\mathbf{y}\|_2^2 + \alpha\|\mathbf{x}\|_{\mathsf{TV}} + \mu\|W\mathbf{x}\|_1 + \beta\sum_{i=1}^{s}\|(\mathbf{z})_{\tilde{g}_i}\|_2 + \frac{\lambda}{2}\|\mathbf{z}-GW\mathbf{x}\|_2^2 \quad (12)$$

Two subproblems:

- $\min_{\mathbf{z}_{\tilde{g}_i}} \beta\|(\mathbf{z})_{\tilde{g}_i}\|_2 + \frac{\lambda}{2}\|\mathbf{z}_{\tilde{g}_i} - (GW\mathbf{x})_{\tilde{g}_i}\|_2^2$ is solved by proximity operator.
- $\min_{\mathbf{x}} \frac{1}{2}\|M\mathbf{x}-\mathbf{y}\|_2^2 + \frac{\lambda}{2}\|\mathbf{z}-GW\mathbf{x}\|_2^2 + \alpha\|\mathbf{x}\|_{\mathsf{TV}} + \mu\|W\mathbf{x}\|_1$ is solved by FISTA
- Proximal operator of $\alpha\|\mathbf{x}\|_{\mathsf{TV}} + \mu\|W\mathbf{x}\|_1$ is solved with an iterative algorithm
- Fast empirical convergence but does not allow parallelization
- No guarantee and does not solve the original problem nor the augmented problem

Solving with the Primal-Dual Framework [2]

Mixture Model

$$\min_{\mathbf{x}} \frac{1}{2}\|\underbrace{M\mathbf{x}-\mathbf{y}}_{\mathbf{x}_0}\|_2^2 + \alpha\|\underbrace{\mathbf{x}}_{\mathbf{x}_1}\|_{\mathsf{TV}} + \mu\|\underbrace{W\mathbf{x}}_{\mathbf{x}_2}\|_1 + \beta\sum_{i=1}^{s}\|(\underbrace{GW\mathbf{x}}_{\mathbf{x}_3})_{\tilde{g}_i}\|_2 \qquad (13)$$

$f_0(\mathbf{x}_0) = \|\mathbf{x}_0\|_2^2$, $f_1(\mathbf{x}_1) = \alpha\|\mathbf{x}_1\|_{\mathsf{TV}}$, $f_2(\mathbf{x}_2) = \mu\|\mathbf{x}_2\|_1$, $f_3(\mathbf{x}_3) = \beta\sum_{i=1}^{s}\|(\mathbf{x}_3)_{\tilde{g}_i}\|_2$

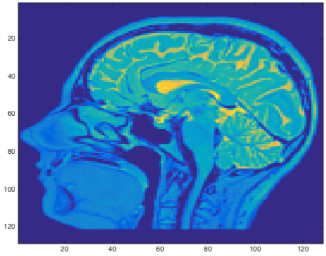
Decomposable form

$$\begin{array}{ll} \min_{\mathbf{x}=[\mathbf{x}_0^T,\ldots,\mathbf{x}_p^T]^T} & f(\mathbf{x}) := \sum_{i=0}^{3} f_i(\mathbf{x}_i) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \qquad (14)$$

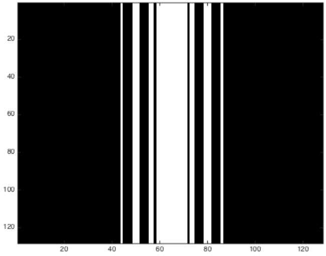
where

$$A = \begin{bmatrix} W & -\boldsymbol{I} & 0 & 0 \\ 0 & G & -\boldsymbol{I} & 0 \\ M & 0 & 0 & -\boldsymbol{I} \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \mathbf{y} \end{bmatrix} \qquad (15)$$

Experimental Setup



Original image



Subsampling map

- $N = 128 \times 128$ MRI brain image sampled via a partial Fourier operator at a subsampling ratio of 0.2
- Note that although we use the same coefficient values for α, β, μ, WaTMRI addresses the augmented problem without the constraint.

Results

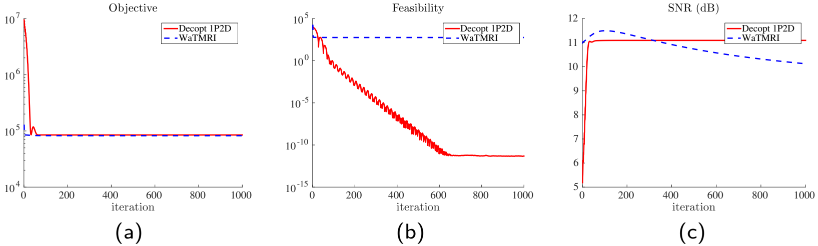


Figure : MRI experiment. (a) Objective function vs iterations. (b) Feasibility gap $\|\mathbf{z} - GW\mathbf{x}\|_2$ vs iterations. (c) Signal-to-noise ratio of the iterates vs iterations.

Outline

Conclusion

- Reliable solver for mixture of regularizers
- Convergence guarantee on both objective and feasibility gap
- Can handle as many regularizers as we want
- Requires only proximal operator computations and parallelizable

References

[1] C. Chen and J. Huang.
Compressive sensing mri with wavelet tree sparsity.
In *Advances in neural information processing systems*, 2012.

[2] Baran Gözcü, Luca Baldassarre, Quoc Tran-Dinh, Cosimo Aprile, and Volkan Cevher.
A primal-dual framework for mixtures of regularizers.
2015.

[3] Yu. Nesterov.
Smooth minimization of non-smooth functions.
Math. Program., Ser. A, 103:127–152, 2005.

[4] Quoc Tran-Dinh and Volkan Cevher.
Constrained convex minimization via model-based excessive gap.
In *Conference of Neural Information Processing Systems (NIPS)*, 2014.

[5] Quoc Tran-Dinh and Volkan Cevher.
A primal-dual algorithmic framework for constrained convex minimization.
Technical report, EPFL, 2014.

1P2D Algorithm

Update the primal-dual sequence $\{\bar{\mathbf{z}}^k\}$

We can design different strategies to update $\{\mathbf{z}^k\}$. For instance:

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\bar{\lambda}^k + \tau_k \lambda^\star_{\beta_k}(\bar{\mathbf{x}}^k) \\ \bar{\mathbf{x}}^{k+1} & := (1-\tau_k)\bar{\mathbf{x}}^k + \tau_k \mathbf{x}^\star_{\gamma_{k+1}}(\hat{\lambda}^k) \\ \bar{\lambda}^{k+1} & := \hat{\lambda}^k + \alpha_k(\mathbf{A}\mathbf{x}^\star_{\gamma_{k+1}}(\hat{\lambda}^k) - \mathbf{b}) \end{cases} \tag{1P2D}$$

where $\alpha_k := \gamma_{k+1}\|\mathbf{A}\|^{-2}$ (Bregman), or $\alpha_k := \gamma_{k+1}$ (augmented Lagrangian).

Update parameters

The parameters β_k and γ_k are updated as ($c_k \in (-1, 1]$ given):

$$\gamma_{k+1} := (1 - c_k\tau_k)\gamma_k \quad \text{and} \quad \beta_{k+1} = (1-\tau_k)\beta_k \tag{16}$$

The parameter τ_k is updated as:

$$a_{k+1} := \left(1 + c_{k+1} + \sqrt{4a_k^2 + (1-c_{k+1})^2}\right)/2, \text{ and } \tau_{k+1} = a_{k+1}^{-1}.$$