Tradeoffs in primal-dual optimization

Volkan Cevher
volkan.cevher@epfl.ch

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

Seminaire Parisien d'Optimisation

ENSCP

[April 13, 2015]

Joint work with

Alp Yurtsever and Quoc Tran Dinh @ LIONS

lions@epfl

MARIE CURIE ACTIONS

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

erc

EPFL

Swiss army knife of convex formulations

A **primal problem** prototype[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\}, \tag{1}$$

- 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 (1) 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 body $\mathcal{C}$ without any fundamental change.

Swiss army knife of convex formulations

A **primal problem** prototype[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\}, \tag{1}$$

- 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 (1) 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

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

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

Swiss army knife of convex formulations

A **primal problem** prototype[1]

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \mathbf{x} \in \mathcal{X} \right\}, \tag{1}$$

- 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 (1) 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

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

Broader context for (1):

- Standard convex optimization formulations: *linear programming, convex quadratic programming, second order cone programming, semidefinite programming and geometric programming*.
- 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 body $\mathcal{C}$ without any fundamental change.

Numerical ϵ-accuracy

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable** unless problem has a closed form solution, which is extremely limited in reality.
- Numerical optimization algorithms result in $\mathbf{x}_\epsilon^\star$ that approximates $\mathbf{x}^\star$ up to a given accuracy ϵ in **some sense**.

- In the sequel, by ϵ-accurate solutions $\mathbf{x}_\epsilon^\star$ of (1), we mean the following

Definition (ϵ-accurate solutions)

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}_\epsilon^\star \in \mathbb{R}^p$ is called an ϵ-solution of (1) if

$$\begin{cases} |f(\mathbf{x}_\epsilon^\star) - f^\star| \leq \epsilon & \text{(objective residual)}, \\ \|\mathbf{A}\mathbf{x}_\epsilon^\star - \mathbf{b}\| \leq \epsilon & \text{(feasibility gap)}, \\ \mathbf{x}_\epsilon^\star \in \mathcal{X} & \text{(exact simple set feasibility)}.^2 \end{cases}$$

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).
- Indeed, ϵ can be different for the objective, feasibility gap, or the iterate residual.

[2] Very often, $\mathcal{X}$ is a "simple set." Hence, requiring $\mathbf{x}_\epsilon^\star \in \mathcal{X}$ is acceptable in practice.*

* **I will absorb $\mathcal{X}$ into the objective f with a so-called indicator function to ease the notation.**

Performance of optimization algorithms

Time-to-reach ϵ

`time-to-reach` ϵ `= number of iterations to reach` ϵ $\times$ `per iteration time`

The **speed** of numerical solutions depends on two factors:

- **Convergence rate** determines the number of iterations needed to obtain an ϵ-optimal solution.
- **Per-iteration time** depends on the information oracles, implementation, and the computational platform.

Finding the fastest algorithm is tricky!

We will discuss basic tradeoffs in primal-dual optimization in the sequel.

Outline

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

$$\begin{cases} 0 & \in \mathbf{A}^T\lambda^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases} \tag{2}$$

(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 (2) 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 (1), 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.

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (1), 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 (1)

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (1), 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 (1)

Efficiency considerations for the dual problem

Subgradient method

1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$$\lambda^{k+1} = \lambda^k + \alpha_k \mathbf{v}^k,$$
where $\mathbf{v}^k \in \partial d(\lambda^k)$ and α_k is the step-size.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$.
Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}}$$

Efficiency considerations for the dual problem

Subgradient method

1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$$\lambda^{k+1} = \lambda^k + \alpha_k \mathbf{v}^k,$$
where $\mathbf{v}^k \in \partial d(\lambda^k)$ and α_k is the step-size.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$.
Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}} \leq \bar{\epsilon}$$

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ $\times$ subgradient calculation

Efficiency considerations for the dual problem

Gradient method

1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$$\lambda^{k+1} = \lambda^k + \frac{1}{L}\nabla d(\lambda^k),$$

where L is the Lipschitz constant.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$.
Then, the subgradient method satisfies

$$\min_{0\leq i\leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}} \leq \bar{\epsilon}$$

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ $\times$ subgradient calculation

GM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$ $\times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the gradient method with step-size $1/L$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{k+4} \leq \bar{\epsilon}.$$

Efficiency considerations for the dual problem

Gradient method

1. Choose $\lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$$\lambda^{k+1} = \lambda^k + \frac{1}{L}\nabla d(\lambda^k),$$

where L is the Lipschitz constant.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$.
Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}} \leq \bar{\epsilon}$$

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ $\times$ subgradient calculation

GM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$ $\times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the gradient method with step-size $1/L$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{k+4} \leq \bar{\epsilon}.$$

This is NOT the best we can do.

There exists a complexity lower-bound

$$d^\star - d(\lambda^k) \geq \frac{3LR^2}{32(k+1)^2}, \forall d(\lambda) \in \mathcal{F}_L,$$

for any iterative method based only on function and gradient evaluations.

Efficiency considerations for the dual problem

Accelerated gradient method

1. Choose $\mathbf{u}^0 = \lambda^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:

$$\lambda^k = \mathbf{u}^k + \frac{1}{L}\nabla d(\mathbf{u}^k),$$
$$\mathbf{u}^{k+1} = \lambda^k + \rho_k(\lambda^k - \lambda^{k-1}),$$

where L is the Lipschitz constant, and ρ_k is a momentum parameter.

Subgradient method for the dual

Assume that the following conditions

1. $\|\mathbf{v}\|_2 \leq G$ for all $\mathbf{v} \in \partial d(\lambda)$, $\lambda \in \mathbb{R}^n$.
2. $\|\lambda^0 - \lambda^\star\|_2 \leq R$

Let the step-size be chosen as $\alpha_k = \frac{R}{G\sqrt{k}}$.
Then, the subgradient method satisfies

$$\min_{0 \leq i \leq k} d^\star - d(\lambda^i) \leq \frac{RG}{\sqrt{k}} \leq \bar{\epsilon}$$

SGM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ $\times$ subgradient calculation

GM: $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$ $\times$ gradient calculation

AGM: $\mathcal{O}\left(\frac{1}{\sqrt{\bar{\epsilon}}}\right)$ $\times$ gradient calculation

Impact of smoothness

(Lipschitz gradient) $d(\lambda)$ has Lipschitz continuous gradient iff

$$\|\nabla d(\lambda) - \nabla d(\eta)\|_2 \leq L\|\lambda - \eta\|_2$$

for all $\lambda, \eta \in \mathrm{dom}(d)$ and we indicate this structure as $d(\lambda) \in \mathcal{F}_L$.

For all $d(\lambda) \in \mathcal{F}_L$, the accelerated gradient method with momentum $\rho_k = \frac{k+1}{k+3}$ obeys

$$d^\star - d(\lambda^k) \leq \frac{2LR^2}{(k+2)^2} \leq \bar{\epsilon}$$

This is NEARLY the best we can do.

There exists a complexity lower-bound

$$d^\star - d(\lambda^k) \geq \frac{3LR^2}{32(k+1)^2}, \forall d(\lambda) \in \mathcal{F}_L,$$

for any iterative method based only on function and gradient evaluations.

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.

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

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 $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ 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.

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

Nesterov's smoother [9]

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

Computational efficiency: The key role of the prox-operator

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:

- single valued & non-expansive.
- 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.

Computational efficiency: The key role of the prox-operator

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:

- single valued & non-expansive.
- 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.

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

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

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 (1)

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 (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 (1)

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 (1).

Measuring progress via the gap function

Acssociated with MVIP, we can define a **gap function** to measure our progress

$$G(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{X} \times \mathbb{R}^n} \left\{ f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) \right\}. \tag{3}$$

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 (1).
- 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 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 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 new technique: Model-based gap reduction MGR (cf., [12])

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

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

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

Key estimates [12, 13]

As a consequence of MGR, we can obtain

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

where $D_{\Lambda^\star} := \min\{\|\lambda^\star\| \ : \ \lambda^\star \in \Lambda^\star\}$ the **norm** of the minimum norm dual solution.

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

Key estimates [12, 13]

As a consequence of MGR, we can obtain

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

where $D_{\Lambda^\star} := \min\{\|\lambda^\star\| \ : \ \lambda^\star \in \Lambda^\star\}$ the **norm** of the minimum norm dual solution.

An uncertainty relation via MGR

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

$$\boxed{\gamma_k \beta_k \geq \frac{\tau_k^2}{1-\tau_k^2}\|\mathbf{A}\|^2}$$

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

Recall: Uncertainty relation

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

$$\boxed{\gamma_k \beta_k \geq \frac{\tau_k^2}{1-\tau_k^2} \|\mathbf{A}\|^2}$$

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

Theorem [12, 13]

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 $\beta_k = \mathcal{O}\left(\frac{1}{k}\right)$:

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

Accelerated gradient method (expanded)

The standard scheme ([11])

The accelerated scheme for minimizing $g \in \mathcal{F}_L^{1,1}$ consists of three main steps:

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k \lambda_k^* \\ \lambda^{k+1} & := \hat{\lambda}^k - \frac{1}{L_g}\nabla g(\hat{\lambda}^k) \\ \lambda_{k+1}^* & := \lambda_k^* - \frac{1}{\tau_k}(\hat{\lambda}^k - \lambda^{k+1}). \end{cases} \tag{4}$$

Here, L_g is the Lipschitz constant of ∇g and $\tau_k \in (0,1)$ is a given momentum term.

Accelerated gradient scheme for the smoothed dual problem

Recall the smoothed dual function d_γ. The AGM for this problem can be written as

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k \lambda_k^* \\ \lambda^{k+1} & := \hat{\lambda}^k + \frac{\gamma}{\|\mathbf{A}\|^2}(\mathbf{A}\mathbf{x}_\gamma^*(\hat{\lambda}^k) - \mathbf{b}) \\ \lambda_{k+1}^* & := \lambda_k^* - \frac{1}{\tau_k}(\hat{\lambda}^k - \lambda^{k+1}). \end{cases} \tag{5}$$

Here, $L_d := \frac{\|\mathbf{A}\|^2}{\gamma}$ and $\nabla d_\gamma(\hat{\lambda}^k) = \mathbf{A}\mathbf{x}_\gamma^*(\hat{\lambda}^k) - \mathbf{b}$.

Our primal-dual scheme

The primal-dual scheme

Our approach is fundamentally the same as the accelerated gradient method:

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k \tilde{\lambda}^k \\ \lambda^{k+1} & := \hat{\lambda}^k + \frac{\gamma_{k+1}}{\|\mathbf{A}\|^2}(\mathbf{A}\mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k) - \mathbf{b}) \\ \bar{\mathbf{x}}^{k+1} & := (1-\tau_k)\bar{\mathbf{x}}^k + \tau_k \mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k) \\ \tilde{\lambda}^{k+1} & := \frac{1}{\beta_{k+1}}(\mathbf{A}\bar{\mathbf{x}}^{k+1} - \mathbf{b}). \end{cases} \tag{6}$$

Both smoothing parameters γ and β are updated at each iteration.

The correspondance between (5) and (6)

The last step of (6) (vs. (5)) is split into two steps to save a matrix multiplication:

$$\frac{1}{\beta_{k+1}}(\mathbf{A}\bar{\mathbf{x}}^{k+1} - \mathbf{b}) = \frac{1}{\beta_k}(\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}) + \frac{\gamma_{k+1}}{\tau_k\|\mathbf{A}\|^2}(\mathbf{A}\mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k) - \mathbf{b}).$$

Using $\bar{\mathbf{x}}^{k+1} := (1-\tau_k)\bar{\mathbf{x}}^k + \tau_k \mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k)$ we can show that

$$\begin{cases} \beta_{k+1} = (1-\tau_k)\beta_k \\ \beta_{k+1}\gamma_{k+1} = \tau_k^2\|\mathbf{A}\|^2. \end{cases}$$

The general constraint case

Handling a cone constraint $\mathbf{Ax} - \mathbf{b} \in \mathcal{K}$

Two steps need to be changed:

$$\begin{cases} \lambda^{k+1} & := \mathrm{proj}_{\mathcal{K}^*} \left(\hat{\lambda}^k + \frac{\gamma}{\|\mathbf{A}\|^2} (\mathbf{A}\mathbf{x}^*_\gamma(\hat{\lambda}^k) - \mathbf{b}) \right) \\ \lambda^*_{k+1} & := \arg\max_{\lambda \in \mathcal{K}^*} \left\{ \langle \mathbf{A}\bar{\mathbf{x}}^{k+1} - \mathbf{b}, \lambda \rangle - \beta_{k+1} p(\lambda) \right\}. \end{cases}$$

Here, $\mathcal{K}^*$ is the dual cone of $\mathcal{K}$, $\mathrm{proj}_{\mathcal{K}^*}$ is the projection onto $\mathcal{K}^*$, and p is a chosen proximity function.

Augmented Lagrangian idea: From $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{1}{\sqrt{\epsilon}}\right)$

When else can the dual function have Lipschitz gradient?

We can render $f(\mathbf{x})$ γ-strongly convex only on $\mathbf{Ax}=\mathbf{b}$ by augmenting it:

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

Denoting $\mathbf{x}_\gamma^*(\lambda)$ as the augmented Lagrangian problem solution, we observe:

- The dual function $d(\lambda)$ is $\frac{1}{\gamma}$-Lipschitz gradient.
- The gradient is $\nabla d_\gamma(\lambda) = \mathbf{A}\mathbf{x}_\gamma^*(\lambda) - \mathbf{b}$.
- The optimal dual objective value does not change: $d^\star = d_\gamma^\star$!!!

Augmented Lagrangian idea: From $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ to $\mathcal{O}\left(\frac{1}{\sqrt{\bar{\epsilon}}}\right)$

When else can the dual function have Lipschitz gradient?

We can render $f(\mathbf{x})$ γ-strongly convex only on $\mathbf{Ax}=\mathbf{b}$ by augmenting it:

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

Denoting $\mathbf{x}_\gamma^*(\lambda)$ as the augmented Lagrangian problem solution, we observe:

- The dual function $d(\lambda)$ is $\frac{1}{\gamma}$-Lipschitz gradient.
- The gradient is $\nabla d_\gamma(\lambda) = \mathbf{A}\mathbf{x}_\gamma^*(\lambda) - \mathbf{b}$.
- The optimal dual objective value does not change: $d^\star = d_\gamma^\star$!!!

Augmented Lagrangian smoother

We augment the Lagrange subproblem so that the dual is smooth!

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

1. λ^k of AGD on $d_\gamma(\lambda)$ has $d^\star - d_\gamma(\lambda^k) \leq \frac{2R^2}{\gamma(k+2)^2}$.
2. We obtain $d^\star - d_\gamma(\lambda^k) \leq \bar{\epsilon}$ with $k = \mathcal{O}\left(\frac{R}{\sqrt{\gamma\bar{\epsilon}}}\right)$.

Augmented Lagrangian idea: The tradeoffs

Key estimates

As a consequence of MGR, we can obtain

$$-\tfrac{\gamma}{2}\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\|^2-\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\|D_{\Lambda^\star}\leq \quad f(\mathbf{x}^k)-f^\star \quad \leq 0$$
$$\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\| \quad \leq \tfrac{2D_{\Lambda^\star}\beta_k}{1-\gamma\beta_k}.$$

where $D_{\Lambda^\star}:=\min\{\|\lambda^\star\| \ : \ \lambda^\star\in\Lambda^\star\}$ the **norm** of the minimum norm dual solution.

An uncertainty relation via MGR

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

$$\gamma\beta_{k+1}\geq\tau_k^2.$$

Here, we update β_k as $\beta_{k+1}=(1-\tau_k)\beta_k$. Then $\beta_k=\Omega(\tau_k^2)$.
Note that $\tau_k^2=\Omega\left(\frac{1}{k^2}\right)$ due to Nesterov's lowerbound.

- The rate of β_k controls the primal residual: $|f(\mathbf{x}^k)-f^\star|\leq\mathcal{O}\left(\beta_k\right)$
- The rate of β_k controls the feasibility: $\|\mathbf{A}\mathbf{x}^k-\mathbf{b}\|_2\leq\mathcal{O}\left(\beta_k\right)$
- They can be simultaneously $\mathcal{O}\left(\frac{1}{k^2}\right)$!

Augmented Lagrangian idea: The tradeoffs

Our augmented primal-dual scheme

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k\tilde{\lambda}^k \\ \lambda^{k+1} & := \hat{\lambda}^k + \gamma(\mathbf{A}\mathbf{x}^*_\gamma(\hat{\lambda}^k) - \mathbf{b}) \\ \bar{\mathbf{x}}^{k+1} & := (1-\tau_k)\bar{\mathbf{x}}^k + \tau_k\mathbf{x}^*_\gamma(\hat{\lambda}^k) \\ \tilde{\lambda}^{k+1} & := \frac{1}{\beta_{k+1}}(\mathbf{A}\bar{\mathbf{x}}^{k+1} - \mathbf{b}). \end{cases} \tag{7}$$

The update rule for parameters:

$$\beta_{k+1} := (1-\tau_k)\beta_k \quad \text{and} \quad \tau_{k+1} = 0.5\tau_k\left[\sqrt{\tau_k^2+4} - \tau_k\right].$$

Theorem (convergence guarantee)

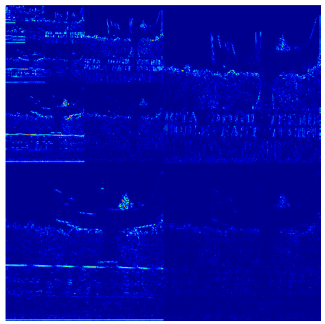
The sequence $\{\bar{\mathbf{z}}^k\}$ generated by (7) satisfies:

$$\begin{aligned} -\tfrac{\gamma}{2}\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|^2 - \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|D_{\Lambda^\star} \leq \quad & f(\mathbf{x}^k) - f^\star \quad \leq 0 \\ & \|\mathbf{A}\mathbf{x}^k - \mathbf{b}\| \quad \leq \tfrac{8D_{\Lambda^\star}}{\gamma(k+1)^2}. \end{aligned}$$

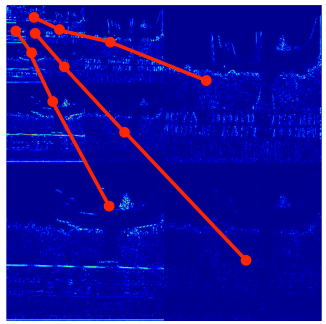
The worst-case iteration complexity: $\mathcal{O}\left(\sqrt{\frac{D_{\Lambda^\star}}{\gamma\epsilon}}\right)$.

- We can increase γ to obtain faster convergence
- However, it becomes more difficult to compute $\mathbf{x}^*_\gamma(\hat{\lambda}^k)$!

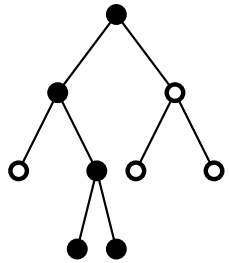
Tree sparsity [7, 4, 2, 15]



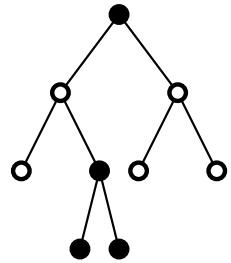
Wavelet coefficients



Wavelet tree



Valid selection of nodes



Invalid selection of nodes

Structure: *We seek the sparsest signal with a rooted connected subtree support.*

Optimization formulation (TU-relax [5])

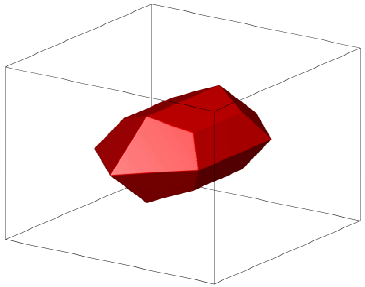
$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) := \sum\nolimits_{\mathcal{G}_i\in\mathfrak{G}} \|\mathbf{x}_{\mathcal{G}_i}\|_\infty \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{8}$$

This problem possesses two key structures: decomposability and tractable proximity.

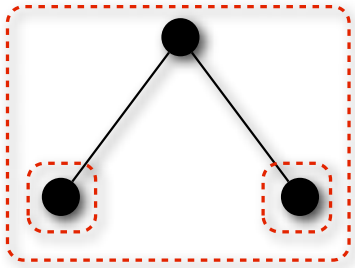
When $g = p$ and $\mathcal{G}_i = \{i\}$, (8) reduces to the well-known basis pursuit (BP):

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{9}$$

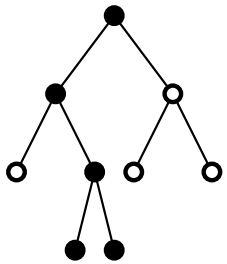
Tree sparsity [7, 4, 2, 15]



$f(\mathbf{x})$-ball



$\mathfrak{G} = \{\{1,2,3\},\{2\},\{3\}\}$



valid selection of nodes

Structure: *We seek the sparsest signal with a rooted connected subtree support.*

Optimization formulation (TU-relax [5])

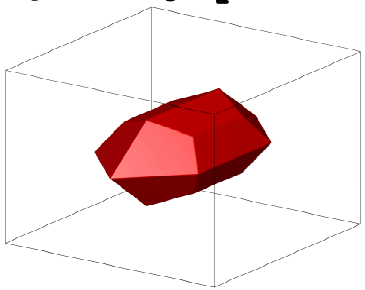
$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) := \sum_{\mathcal{G}_i\in\mathfrak{G}} \|\mathbf{x}_{\mathcal{G}_i}\|_\infty \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{8}$$

This problem possesses two key structures: decomposability and tractable proximity.

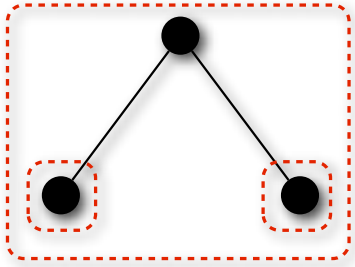
When $g = p$ and $\mathcal{G}_i = \{i\}$, (8) reduces to the well-known basis pursuit (BP):

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{9}$$

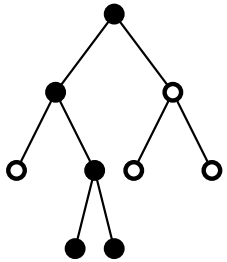
Tree sparsity [7, 4, 2, 15]



$f(\mathbf{x})$-ball



$\mathfrak{G} = \{\{1,2,3\},\{2\},\{3\}\}$



valid selection of nodes

Structure: *We seek the sparsest signal with a rooted connected subtree support.*

Optimization formulation (TU-relax [5])

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) := \sum_{\mathcal{G}_i\in\mathfrak{G}} \|\mathbf{x}_{\mathcal{G}_i}\|_\infty + \alpha\|\mathbf{\Psi x}\|_{\mathsf{TV}} \\ \text{s.t.} & \mathbf{Ax} = \mathbf{b}. \end{array} \tag{8}$$

This problem possesses two key structures: decomposability and tractable proximity.

When $g = p$ and $\mathcal{G}_i = \{i\}$, (8) reduces to the well-known basis pursuit (BP):

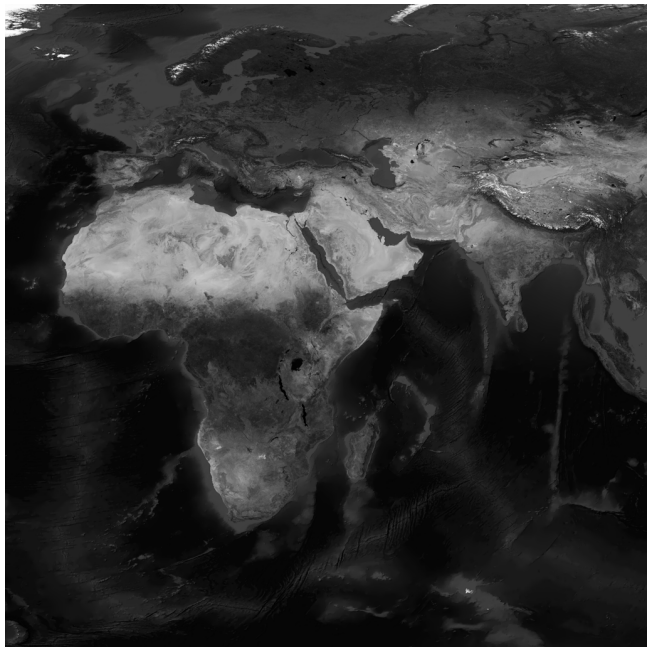
$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}\|_1 + \alpha\|\mathbf{\Psi x}\|_{\mathsf{TV}} \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{b}. \tag{9}$$

Adding additional regularizers to BP does not pose any difficulty

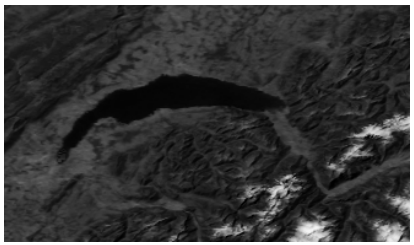
- $\mathbf{\Psi}$ is the wavelet transform and α is a regularization parameter

Tree sparsity example: 1:100-compressive sensing [12, 1]

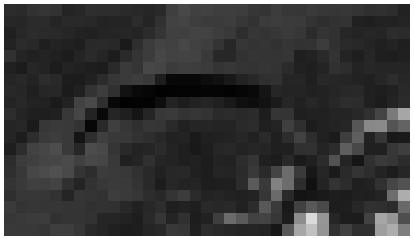
World [1Gpix]



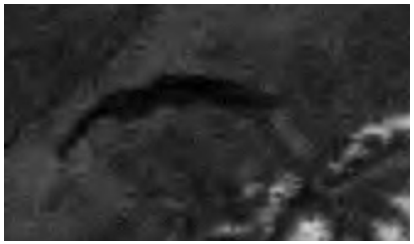
Lac Léman



World [10Mpix]

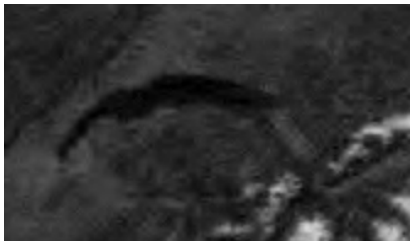


sparse



PNSR $= 31.83$db

TU-relax



PNSR $= 32.48$db

Augmented Lagrangian method
Iterations: 113
PD gap: 1e−8
Applications of $(\mathbf{A}, \mathbf{A}^T)$: $(684, 570)$
Time: < 4days

Tree sparsity example: TV & TU-relax 1:15-compression [12, 1]

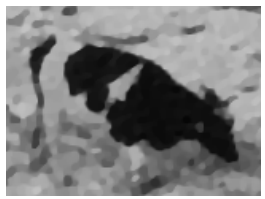
Original tiff image [2048×2048]

Original

BP

TU-relax

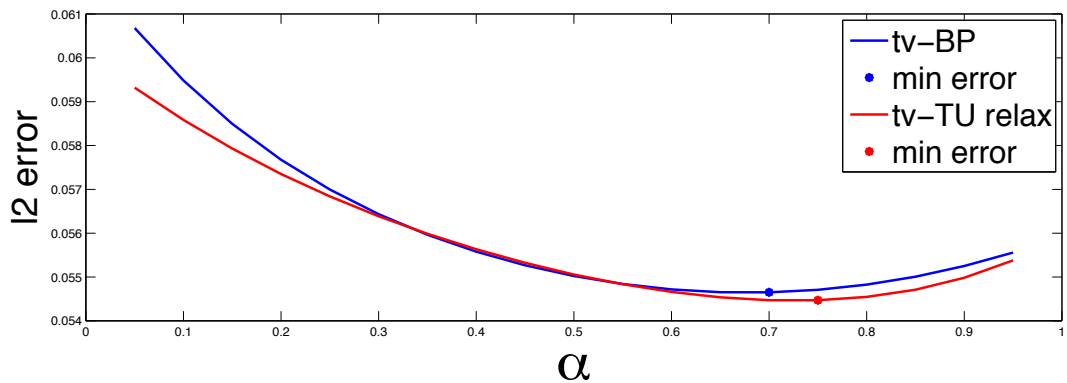
TV



TV with BP

TV with TU-relax

Regularization:



Outline

Revisiting the prox-operator

Prox-operator helps us process nonsmooth terms "efficiently"

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

Key properties:

- single valued & non-expansive.
- 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.

Revisiting the prox-operator

Prox-operator helps us process nonsmooth terms "efficiently"

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

Key properties:

- single valued & non-expansive.
- 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.

Not all nonsmooth functions are proximal-friendly!

If $g(\mathbf{z}) = \|\mathbf{z}\|_\star$ (i.e., the nuclear norm of $\mathbf{z}$) then the prox-operator may require a full singular value decomposition.

We will discuss how to avoid the prox-operator whenever it is expensive!

Example: Frank-Wolfe's method

Problem setting

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})} \tag{10}$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).
- Note that $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}$ is missing from our prototype problem

Frank-Wolfe's method (see [6] for a review)

Conditional gradient method (CGA)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$.

2. For $k = 0, 1, \cdots$, perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \arg\min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T \mathbf{x}, \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$ is a given relaxation parameter.

Example: Frank-Wolfe's method

Problem setting

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})} \tag{10}$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).
- Note that $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}$ is missing from our prototype problem

Frank-Wolfe's method (see [6] for a review)

Conditional gradient method (CGA)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$.

2. For $k = 0, 1, \cdots$, perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \arg\min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T \mathbf{x}, \ (*) \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k \hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$ is a given relaxation parameter.

When $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^{n\times p} : \|\mathbf{x}\|_\star \leq 1\}$, $(*)$ corresponds to rank-1 updates!

Towards Fenchel-type operators

Generalized sharp operators [14]

We define the (generalized) sharp operator of a convex function g over $\mathcal{X}$ as follows

$$\boxed{[\mathbf{x}]^\sharp_{\mathcal{X},g} := \operatorname*{argmax}_{\mathbf{z}\in\mathcal{X}} \{\langle \mathbf{x}, \mathbf{z}\rangle - g(\mathbf{z})\}\,.}$$

Important special cases:

1. If $g = 0$, then we obtain the so-called linear minimization oracle.
2. If $\mathcal{X} = \operatorname{dom}(g)$, then $[\mathbf{x}]^\sharp_g = \nabla g^*(\mathbf{x})$, where g^* is the Fenchel conjugate of g.

Example (Nuclear norm)

Consider $g(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|^2_\star$ and $\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^{n\times p} : \|\mathbf{x}\|_\star \leq 1\}$. Let $\mathbf{u}$ and $\mathbf{v}$ be the left and right principal singular vectors of $\mathbf{x}$ respectively. Then,

$$\mathbf{u}\mathbf{v}^T \in [\mathbf{x}]^\sharp_{\mathcal{X}} := [\mathbf{x}]^\sharp_{\mathcal{X},0}\,, \qquad \|\mathbf{x}\|\mathbf{u}\mathbf{v}^T \in [\mathbf{x}]^\sharp_g := [\mathbf{x}]^\sharp_{\mathbb{R}^{n\times p},g}$$

where $\|\cdot\|$ is the spectral norm. **The computations are essentially the same.**

Revisiting Frank-Wolfe's method

Problem setting

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})} \tag{11}$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).
- Note that $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}$ is missing from our prototype problem

Frank-Wolfe's method (see [6] for a review)

Conditional gradient method (CGA)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$.

2. For $k = 0, 1, \cdots$, perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \arg\min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T\mathbf{x} \quad \equiv [\nabla f(\mathbf{x}^k)]_{\mathcal{X}}^\sharp, \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k\hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$ is a given relaxation parameter.

Generalized conditional gradient method replaces the indicator function $\delta_{\mathcal{X}}$ with g:

$$\hat{\mathbf{x}}^k := \arg\min_{\mathbf{x}}\{g(\mathbf{x}) + \nabla f(\mathbf{x}^k)^T\mathbf{x}\} = [\nabla f(\mathbf{x}^k)]_g^\sharp.$$

Revisiting Frank-Wolfe's method

Problem setting

$$\boxed{f^\star := \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x})} \tag{11}$$

Assumptions

- $\mathcal{X}$ is nonempty, convex, closed and bounded.
- $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ (i.e., convex with Lipschitz gradient).
- **We will handle $\mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K}$ and nonsmooth $f(\mathbf{x})$ in the sequel!**

Frank-Wolfe's method (see [6] for a review)

Conditional gradient method (CGA)

1. Choose $\mathbf{x}^0 \in \mathcal{X}$.

2. For $k = 0, 1, \cdots$, perform:

$$\begin{cases} \hat{\mathbf{x}}^k & := \arg\min_{\mathbf{x}\in\mathcal{X}} \nabla f(\mathbf{x}^k)^T\mathbf{x} \quad \equiv [\nabla f(\mathbf{x}^k)]_{\mathcal{X}}^\sharp, \\ \mathbf{x}^{k+1} & := (1-\gamma_k)\mathbf{x}^k + \gamma_k\hat{\mathbf{x}}^k, \end{cases}$$

where $\gamma_k := \frac{2}{k+2}$ is a given relaxation parameter.

Generalized conditional gradient method replaces the indicator function $\delta_{\mathcal{X}}$ with g:

$$\hat{\mathbf{x}}^k := \arg\min_{\mathbf{x}}\{g(\mathbf{x}) + \nabla f(\mathbf{x}^k)^T\mathbf{x}\} = [\nabla f(\mathbf{x}^k)]_g^\sharp.$$

Exploring the smoothness of the dual function in depth

Definition (Hölder continuous gradients [8])

Let us consider the following unconstrained setup

$$\min_{\mathbf{x}\in\mathbb{R}^p} g(\mathbf{x}).$$

A convex function g has Hölder continuous subgradients of degree ν if there are nonnegative real constants ν and M_ν that satisfy:

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\|_* \leq M_\nu \|\mathbf{x}-\mathbf{y}\|^\nu$$

where $\nabla g(\lambda)$ is a (sub)gradient of g.

Highlights:

1. $\nu = 1$ is the Lipschitz continuous gradients case where $L = M_\nu$.
2. $\nu = 0$ is the bounded gradient assumption (recall the subgradient method).
3. Iteration lowerbound for the Hölder class: $\mathcal{O}\left(\left(\frac{M_\nu \|\mathbf{x}^0 - \mathbf{x}^\star\|^{1+\nu}}{\epsilon} \right)^{\frac{2}{1+3\nu}} \right)$.
4. The condition also ensures a basic surrogate

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y}-\mathbf{x} \rangle + \frac{M_\nu}{1+\nu}\|\mathbf{x}-\mathbf{y}\|^{1+\nu}$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$.

Nesterov's universal gradient methods

Nesterov's universal gradient methods [10]

In practice, smoothness parameters ν and M_ν are usually not known. Nesterov's algorithms adapt to the unknown ν via an appropriate line-search strategy:

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{M}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{\delta}{2}.$$

where inexactness parameter $\delta > 0$ depends only on the desired final accuracy.

They are universal since they ensure the best possible rate of convergence for each ν.

[3] PGM in [10] uses the Bregman / prox setup.

Nesterov's universal gradient methods

Nesterov's universal gradient methods [10]

In practice, smoothness parameters ν and M_ν are usually not known. Nesterov's algorithms adapt to the unknown ν via an appropriate line-search strategy:

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{M}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{\delta}{2}.$$

where inexactness parameter $\delta > 0$ depends only on the desired final accuracy.

They are universal since they ensure the best possible rate of convergence for each ν.

Universal primal gradient method (PGM)[3]

1. Choose $\mathbf{x}^0 \in \mathcal{X}$, $M_{-1} > 0$ and accuracy $\epsilon > 0$.

2. For $k = 0, 1, \cdots$, perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - M_k^{-1} \nabla g(\mathbf{x}^k)$$

where we use line-search to find $M_k \geq 0.5 M_{k-1}$ that satisfies:

$$g(\mathbf{x}^{k+1}) \leq g(\mathbf{x}^k) + \langle \nabla g(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{M_k}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 + \frac{\epsilon}{2}$$

[3] PGM in [10] uses the Bregman / prox setup.

Nesterov's universal gradient methods

Nesterov's universal gradient methods [10]

In practice, smoothness parameters ν and M_ν are usually not known. Nesterov's algorithms adapt to the unknown ν via an appropriate line-search strategy:

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{M}{2} \|\mathbf{x} - \mathbf{y}\|^2 + \frac{\delta}{2}.$$

where inexactness parameter $\delta > 0$ depends only on the desired final accuracy.

They are universal since they ensure the best possible rate of convergence for each ν.

Universal primal gradient method (PGM)[3]

1. Choose $\mathbf{x}^0 \in \mathcal{X}$, $M_{-1} > 0$ and accuracy $\epsilon > 0$.

2. For $k = 0, 1, \cdots$, perform:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - M_k^{-1} \nabla g(\mathbf{x}^k)$$

where we use line-search to find $M_k \geq 0.5 M_{k-1}$ that satisfies:

$$g(\mathbf{x}^{k+1}) \leq g(\mathbf{x}^k) + \langle \nabla g(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k \rangle + \frac{M_k}{2} \|\mathbf{x}^k - \mathbf{x}^{k+1}\|^2 + \frac{\epsilon}{2}$$

Yes, there is an accelerated version.

[3] PGM in [10] uses the Bregman / prox setup.

Universal primal-dual decomposition methods

Our strategy: Hölder smoothness in the dual

Instead of smoothing, we assume that the dual function d is Hölder continuous for some $\nu \in [0,1]$:

$$M_\nu(d) := \sup_{\lambda \neq \eta} \frac{\|\nabla d(\lambda) - \nabla d(\eta)\|_*}{\|\lambda - \eta\|^\nu}, \qquad M_d^* := \inf_{0 \leq \nu \leq 1} M_\nu(d) < +\infty.$$

We will solve the dual problem by a new FISTA version [3] of Nesterov's universal gradient algorithm [10] and develop new primal strategies to approximate $\mathbf{x}^\star$.

Is this assumption reasonable?

Consider two special cases:

- if $\mathcal{X}$ is bounded and d is subdifferentiable, then ∇d is also bounded.
- if f is uniformly convex with convexity parameter $\mu_f > 0$ and degree $q \geq 2$, i.e.,

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \mu_f \|\mathbf{x} - \mathbf{y}\|^q$$

for any $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, then ∇d satisfies Hölder condition with $\nu = \frac{1}{q-1}$ and $M_\nu = \left(\mu_f^{-1} \|\mathbf{A}\|^2\right)^{\frac{1}{q-1}}$.

Our universal primal-dual decomposition methods: The dual steps

Dual steps ([14])

- The universal dual gradient step:

$$\lambda^{k+1} := \lambda^k + \frac{1}{M_k}\nabla d(\lambda^k) = \lambda_k + \frac{1}{M_k}\left(\mathbf{A}\mathbf{x}^*(\lambda^k) - \mathbf{b}\right),$$

  where $\mathbf{x}^*(\lambda^k)$ is computed via the sharp operator:

$$\mathbf{x}^*(\lambda^k) := \arg\min_{\mathbf{x}\in\mathbb{R}^p}\left\{f(\mathbf{x}) + \langle \mathbf{A}^T\lambda^k, \mathbf{x}\rangle\right\} \equiv \left[-\mathbf{A}^T\lambda^k\right]_f^\sharp.$$

- The universal dual accelerated gradient step:

$$\begin{cases} t_k & := 0.5\left(1+\sqrt{1+4t_{k-1}^2}\right) \\ \hat{\lambda}^k & := \lambda^k + \frac{t_{k-1}-1}{t_k}\left(\lambda^k - \hat{\lambda}^{k-1}\right) \\ \lambda^{k+1} & := \hat{\lambda}^k + \frac{1}{M_k}\left(\mathbf{A}\mathbf{x}^*(\hat{\lambda}^k) - \mathbf{b}\right). \end{cases}$$

Line-search condition

The local smoothness constant M_k is computed via a line-search procedure:

$$d(\lambda^{k+1}) \geq d(\lambda^k) + \langle \nabla d(\lambda^k), \lambda^{k+1} - \lambda^k\rangle - \frac{M_k}{2}\|\lambda^{k+1} - \lambda^k\|^2 - \frac{\delta_k}{2}.$$

- $\delta_k = \epsilon$ for our universal dual gradient method
- $\delta_k = \epsilon/t_k$ for our universal dual accelerated gradient method

On the line-search

Number of line-search iterations

- Each line-search step costs one dual function evaluation.
- **(Acc)UniProx** requires (1)2 line-search steps per iteration on the average.
- In many cases, we can avoid the search step and find the step-size in one shot by solving an analytic equation obtained by using a proper bound on $d(\lambda^{k+1})$.

Example (Nuclear norm)

Consider $f := \frac{1}{2}\|\mathbf{x}\|_\star^2$ with the linear constraint $\mathbf{A}(\mathbf{x}) = \mathbf{b}$, which leads to the dual function $d(\lambda) = -\frac{1}{2}\|\mathbf{A}^T(\lambda)\|^2 - \langle \lambda, \mathbf{b}\rangle$. Using triangular inequality, we get

$$U(M_k) := d(\lambda^k) - \frac{\alpha_k^2}{2}\|\mathbf{A}^T(\nabla d(\lambda^k))\|^2 - \alpha_k\left[\|\mathbf{A}^T(\nabla d(\lambda^k))\|\|\mathbf{A}^T(\lambda^k)\| + \langle \nabla d(\lambda^k), \mathbf{b}\rangle\right]$$
$$\leq d(\lambda^k + \frac{1}{M_k}\nabla d(\lambda^k)) = d(\lambda^{k+1}).$$

We can solve the following second order equation

$$U(M_k) = d(\lambda^k) + \frac{\alpha_k}{2}\left\|\nabla d(\lambda^k)\right\|^2 - \frac{\delta_k}{2}$$

to find the step size $\alpha_k := \frac{1}{M_k}$ which guarantees the line-search condition.

The primal steps and the worst-case complexity

Primal steps - averaging steps

- The universal primal gradient step:

$$\textbf{(UniProx):} \qquad \bar{\mathbf{x}}^k := \left(\sum_{i=0}^{k} \frac{1}{M_i}\right)^{-1} \sum_{i=0}^{k} \frac{1}{M_i} \mathbf{x}^*(\lambda^i).$$

- The universal primal accelerated gradient step:

$$\textbf{(AccUniProx):} \qquad \bar{\mathbf{x}}^k := \left(\sum_{i=0}^{k} \frac{t_i}{M_i}\right)^{-1} \sum_{i=0}^{k} \frac{t_i}{M_i} \mathbf{x}^*(\lambda^i).$$

The worst-case complextity

To achieve $\bar{\mathbf{x}}^k$ such that $|f(\bar{\mathbf{x}}^k) - f^\star| \leq \epsilon$ and $\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| \leq \epsilon$ is:

$$\begin{cases} \text{For } \textbf{(UniProx):} & \mathcal{O}\left(D_{\Lambda^\star}^2 \inf\limits_{0\leq\nu\leq 1} \left(\frac{M_\nu}{\epsilon}\right)^{\frac{2}{1+\nu}}\right), \quad (\text{optimal for } \nu = 0). \\ \text{For } \textbf{(AccUniProx):} & \mathcal{O}\left(D_{\Lambda^\star}^{\frac{2+5\nu}{1+3\nu}} \inf\limits_{0\leq\nu\leq 1} \left(\frac{M_\nu}{\epsilon}\right)^{\frac{2}{1+3\nu}}\right), \quad (\text{optimal for } \nu = 1). \end{cases}$$

Summary of the algorithms and convergence guarantees - I

Universal primal-dual gradient method (UniProx)

Initialization: Choose $\lambda^0 \in \mathbb{R}^n$ and $\epsilon > 0$. Estimate a value $M_{-1} < 2M_\epsilon$.

Iteration: For $k = 0, 1, \cdots$, perform:

1. *Primal step:* $\mathbf{x}^*(\lambda^k) = [-\mathbf{A}^T\lambda^k]_f^\sharp$
2. *Dual gradient:* $\nabla d(\lambda^k) = \mathbf{A}^T\mathbf{x}^*(\lambda^k) - \mathbf{b}$
3. Line-search: Find $M_k \in [0.5M_{k-1}, 2M_\epsilon]$ from line-search condition and:
$$\lambda^{k+1} = \lambda^k + M_k^{-1}\nabla d(\lambda^k)$$
4. *Primal averaging:* $\bar{\mathbf{x}}^k := S_k^{-1}\sum_{j=0}^k M_j^{-1}\mathbf{x}^*(\lambda^j)$ where $S_k = \sum_{j=0}^k M_j^{-1}$.

Theorem [14]

$\bar{\mathbf{x}}^k$ and $\bar{\lambda}^k := S_k^{-1}\sum_{j=0}^k M_j^{-1}\lambda^j$ obtained by **UniProx** satisfy (with $\lambda^0 = 0$):

$$\begin{cases} -\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\|D_{\Lambda^\star} \leq & f(\bar{\mathbf{x}}^k) - f^\star \leq \frac{\epsilon}{2}, \\ & \|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| \leq \frac{4M_\epsilon D_{\Lambda^\star}}{k+1} + \sqrt{\frac{2M_\epsilon\epsilon}{k+1}}, \\ & d^\star - d(\bar{\lambda}^k) \leq \frac{M_\epsilon D_{\Lambda^\star}^2}{k+1} + \frac{\epsilon}{2}. \end{cases}$$

Summary of the algorithms and convergence guarantees - II

Accelerated universal primal-dual gradient method (AccUniProx)

Initialization: Choose $\lambda^0 \in \mathbb{R}^n$, $\epsilon > 0$. Set $t_0 = 1$. Estimate a value $M_{-1} < 2M_\epsilon$.

Iteration: For $k = 0, 1, \cdots$, perform:

1. *Primal step:* $\mathbf{x}^*(\hat{\lambda}^k) = [-\mathbf{A}^T\hat{\lambda}^k]_f^\sharp$,
2. *Dual gradient:* $\nabla d(\hat{\lambda}^k) = \mathbf{A}^T\mathbf{x}^*(\hat{\lambda}^k) - \mathbf{b}$,
3. *Line-search:* Find $M_k \in [M_{k-1}, 2M_\epsilon]$ from line-search condition and:
$$\lambda^{k+1} = \hat{\lambda}^k + M_k^{-1}\nabla d(\hat{\lambda}^k),$$
4. $t_{k+1} = 0.5[1 + \sqrt{1 + 4t_k^2}]$,
5. $\hat{\lambda}_{k+1} = \lambda_{k+1} + \frac{t_k - 1}{t_{k+1}}(\lambda_{k+1} - \lambda_k)$,
6. *Primal averaging:* $\bar{\mathbf{x}}^k := S_k^{-1}\sum_{j=0}^k t_j M_j^{-1}\mathbf{x}^*(\lambda^j)$ where $S_k = \sum_{j=0}^k t_j M_j^{-1}$.

Theorem [14]

$\bar{\mathbf{x}}^k$ and λ^k obtained by **AccUniProx** satisfy (with $\lambda^0 = 0$):

$$\begin{cases} -\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\|D_{\Lambda^\star} \leq & f(\bar{\mathbf{x}}^k) - f^\star \leq \frac{\epsilon}{2}, \\ & \|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| \leq \frac{16M_\epsilon D_{\Lambda^\star}}{(k+2)^{\frac{1+3\nu}{1+\nu}}} + \sqrt{\frac{8M_\epsilon\epsilon}{k+2}}, \\ & d^\star - d(\lambda^k) \leq \frac{4M_\epsilon D_{\Lambda^\star}^2}{(k+1)^{\frac{1+3\nu}{1+\nu}}} + \frac{\epsilon M_\epsilon}{M_0}(k+1)^{\frac{1-\nu}{1+\nu}}. \end{cases}$$

The dual may NOT converge!

The general constraint case

Handling to the constraint $\mathbf{Ax} - \mathbf{b} \in \mathcal{K}$

Dual steps need to be changed:

- The universal dual gradient step:

$$\lambda^{k+1} := \operatorname{prox}_{M_k^{-1} h}\left(\lambda_k + \frac{1}{M_k}\left(\mathbf{A}\mathbf{x}^*(\lambda^k) - \mathbf{b}\right)\right).$$

- The universal dual accelerated gradient step:

$$\begin{cases} t_k & := 0.5\left(1 + \sqrt{1 + 4t_{k-1}^2}\right) \\ \hat{\lambda}^k & := \bar{\lambda}^k + \frac{t_{k-1}-1}{t_k}\left(\bar{\lambda}^k - \hat{\lambda}^{k-1}\right) \\ \lambda^{k+1} & := \operatorname{prox}_{M_k^{-1} h}\left(\hat{\lambda}^k + \frac{1}{M_k}\left(\mathbf{A}\mathbf{x}^*(\hat{\lambda}^k) - \mathbf{b}\right)\right). \end{cases}$$

Here, h is defined by $h(\lambda) := \sup_{\mathbf{r} \in \mathcal{K}} \langle \lambda, \mathbf{r} \rangle$.

Example: Robust matrix completion with $\approx 1:50$ subsampling

Problem formulation

Let $\Omega \subseteq \{1, \cdots, p\} \times \{1, \cdots, q\}$ be a subset of indexes and $\mathbf{M}_\Omega = (\mathbf{M}_{ij})_{(i,j)\in\Omega}$ be the observed entries of a missed matrix $\mathbf{M}$. $\mathcal{P}_\Omega$ is the projection on the subset Ω.

$$f^\star := \min_{\mathbf{X}\in\mathbb{R}^{p\times q}} \left\{ f(\mathbf{X}) := \frac{1}{2}\|\mathbf{X}\|_\star^2 : \|\mathcal{P}_\Omega(\mathbf{X}) - \mathbf{M}_\Omega\|_1 \leq \tau \right\}. \tag{12}$$

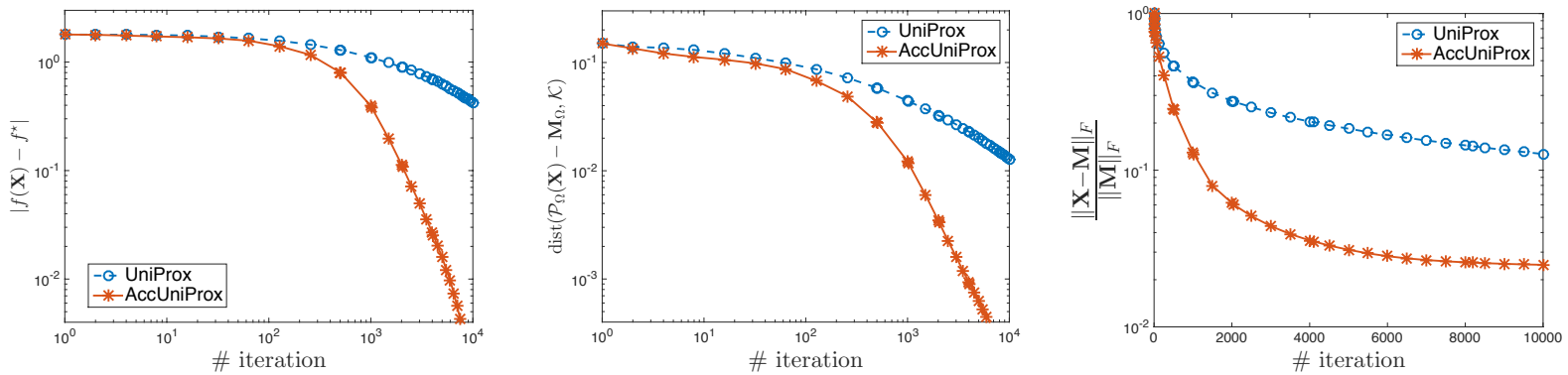


Figure: The performance of UniProx and AccUniProx algorithms for (12).

Setup

- Synthetic data $p = 1000$, $q = 4000$, and rank $r = 6$
- Number of samples $n := |\Omega| = 9 \cdot 10^4$
- Input parameters $\lambda^0 = \mathbf{0}^n$ and $\epsilon = 2 \cdot 10^{-2}$

Example: Nuclear-norm constrained matrix completion - I

Problem formulation

Let $\Omega \subseteq \{1, \cdots, p\} \times \{1, \cdots, q\}$ be a subset of indexes and $\mathbf{M}_\Omega = (\mathbf{M}_{ij})_{(i,j)\in\Omega}$ be the observed entries of a missed matrix $\mathbf{M}$. $\mathcal{P}_\Omega$ is the projection on the subset Ω.

$$f^\star := \min_{\mathbf{X}\in\mathbb{R}^{p\times q}} \left\{ \frac{1}{2}\|\mathcal{P}_\Omega(\mathbf{X}) - \mathbf{M}_\Omega\|_F^2 : \|\mathbf{X}\|_\star \leq \varphi^\star \right\} \tag{13}$$

Setup

- Synthetic data of size $p = 400, q = 2000$ with rank $r = 10$.
- Number of samples $n := |\Omega| = 7.5 \cdot 10^4$.
- $\varphi^\star = \|\mathbf{M}\|_\star$ is assumed to be known.
- Input parameters $\lambda^0 = \mathbf{0}^n$ and $\epsilon = 2 \cdot 10^{-6}$.

Example: Nuclear-norm constrained matrix completion - II

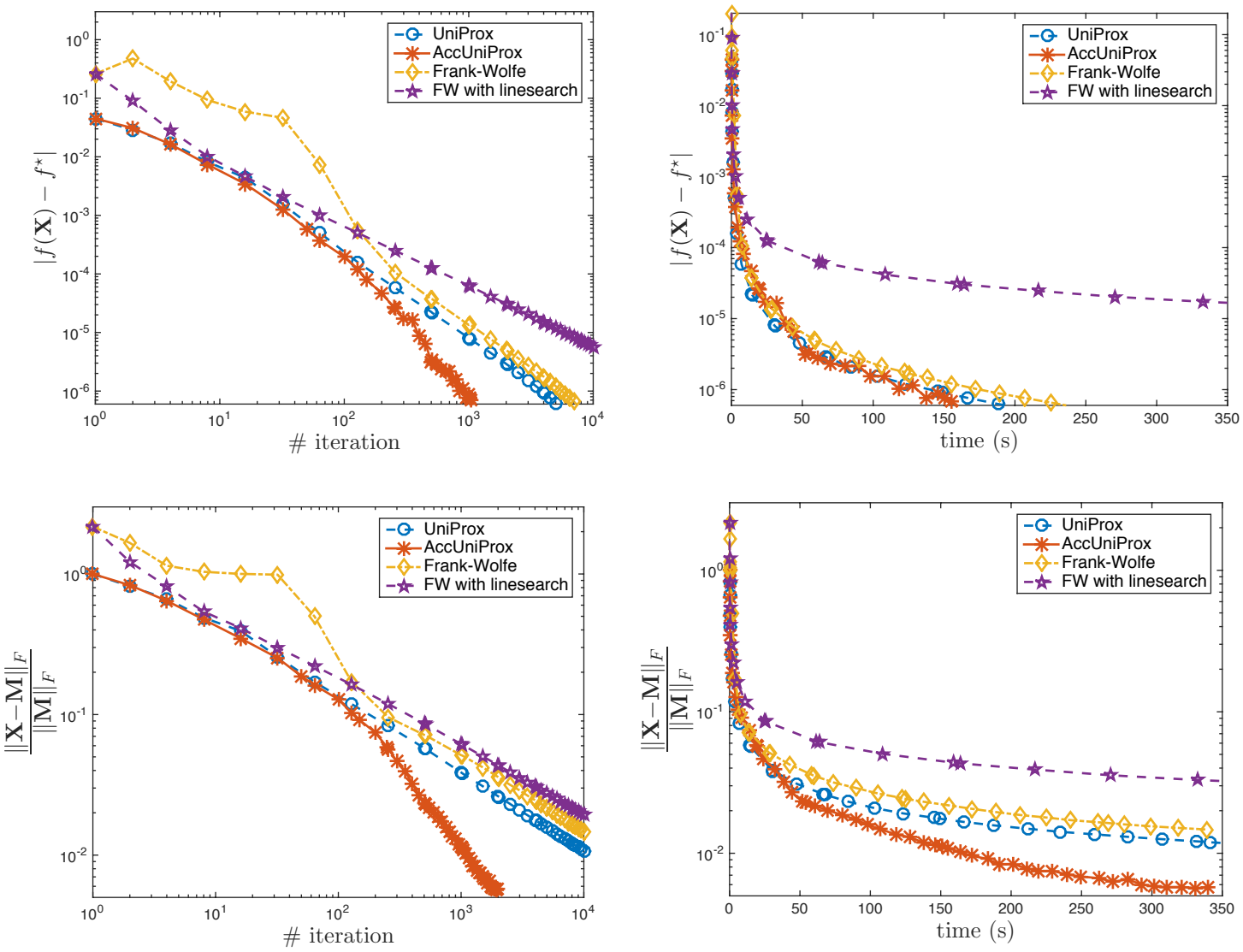


Figure: The performance of (Acc)UniProx and Frank-Wolfe algorithms for (13).

Outline

Conclusions

- Our contributions for the prototype primal problem (1):
  - new proximal primal dual algorithms via the model-based gap reduction technique
  - new universal primal-dual algorithms via sharp operators
  - guarantees on primal optimality and feasibility of the iterates for all
- Uncertainty relations for the proximal way
  - primal optimality of the algorithmic iterates competes with their feasibility
  - clear benefits of the augmented Lagrangian
- The prox-operator vs. the sharp-operator
  - tradeoffs abound (also along the Augmented Lagrangian + ADMM / AMA)
  - time-winner is unclear due to modern data processing systems
  - warm-starts / hybrid strategies ?

References I

[1] Ben Adcock, Anders C. Hansen, Clarice Poon, and Bogdan Roman.
Breaking the coherence barrier: A new theory for compressed sensing.
http://arxiv.org/abs/1302.0561, Feb. 2013.

[2] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.

[3] Amir Beck and Marc Teboulle.
A fast iterative shrinkage-thresholding algorithm for linear inverse problems.
SIAM Journal on Imaging Sciences, 2(1):183–202, 2009.

[4] Marco F. Duarte, Dharmpal Davenport, Mark A. adn Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk.
Single-pixel imaging via compressive sampling.
IEEE Sig. Process. Mag., 25(2):83–91, March 2008.

[5] M. El Halabi and V. Cevher.
A totally unimodular view of structured sparsity.
In *Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2015.

[6] M. Jaggi.
Revisiting frank-wolfe: Projection-free sparse convex optimization.
JMLR W&CP, 28(1):427–435, 2013.

References II

[7] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
Journal of Machine Learning Reasearch, 12:2297–2334, 2011.

[8] A. S. Nemirovsky and D. B. Yudin.
Problem complexity and method efficiency in optimization.
Wiley, New York, 1983.

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

[10] Yu. Nesterov.
Universal gradient methods for convex optimization problems.
Math. Program., pages 1–24, 2014.

[11] Yurii Nesterov.
A method of solving a convex programming problem with convergence rate o (1/k2).
In *Soviet Mathematics Doklady*, volume 27, pages 372–376, 1983.

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

References III

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

[14] Alp Yurtsever, Quoc Tran-Dinh, and Volkan Cevher.
Universal primal-dual proximal-gradient methods.
Technical report, EPFL, 2015.

[15] Peng Zhao, Guilherme Rocha, and Bin Yu.
Grouped and hierarchical model selection through composite absolute penalties.
Department of Statistics, UC Berkeley, Tech. Rep, 703, 2006.