Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 10: Constrained convex minimization I

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

EE-556 (Fall 2017)

License Information for Mathematics of Data Slides

Outline

- Today
  1. Primal-Dual methods
- Next week
  1. Frank-Wolfe method
  2. Universal primal-dual gradient methods
  3. ADMM

Recommended readings

- Quoc Tran-Dinh, Olivier Fercoq and Volkan Cevher, *A Smooth Primal-Dual Optimization Framework for Nonsmooth Composite Convex Minimization.* to appear in SIOPT, 2017.
- Y. Nesterov, *Smooth Minimization of Non-smooth Functions.* Math. Program., Ser. A, 103:127-152, 2005.

Swiss army knife of convex formulations

A **primal problem** prototype

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

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ 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 $\mathbf{x}^\star \in \mathcal{X}$

An example from the sparseland

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \le \kappa, \|\mathbf{x}\|_\infty \le c \right\} \quad \text{(SOCP)}$$

Swiss army knife of convex formulations

A **primal problem** prototype

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

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ 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 $\mathbf{x}^\star \in \mathcal{X}$

An example from the sparseland

$$\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq c \right\} \tag{SOCP}$$

Broad 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, . . .*

Swiss army knife of convex formulations

A **primal problem** prototype

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

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ 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 $\mathbf{x}^\star \in \mathcal{X}$

A key advantage of the unified formulation (1): **Primal-dual methods**

- decentralized collection & storage of data
- cheap per-iteration costs & distributed computation

Broad 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, . . .*

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

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

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

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

Per-iteration time:

first-order methods: Multiplication with $\mathbf{A}$, $\mathbf{A}^T$, and appropriate "prox-operators"

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

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

Per-iteration time:

first-order methods: Multiplication with $\mathbf{A}$, $\mathbf{A}^T$, and appropriate "prox-operators"

A key issue: Number of iterations to reach ϵ

The notion of ϵ-accuracy is elusive in constrained optimization!

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})}$$

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

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

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

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

Our definition of ϵ-accurate solutions [19]

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

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

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \leq \epsilon$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \leq \epsilon \quad !!!$$

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

Our definition of ϵ-accurate solutions [19]

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

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

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).

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

Primal-dual methods for (1):

Plenty ...

- Variants of the **Arrow-Hurwitz's method**:
  - Chambolle-Pock's algorithm [2], and its variants, e.g., He-Yuan's variant [13].
  - Primal-dual Hybrid Gradient (PDHG) method and its variants [9, 11].
  - Proximal-based decomposition (Chen-Teboulle's algorithm) [3].
- Splitting techniques from monotone inclusions:
  - Primal-dual splitting algorithms [1, 4, 21, 5, 6].
  - Three-operator splitting [7].
- Dual splitting techniques:
  - Alternating minimization algorithms (AMA) [10, 21].
  - Alternating direction methods of multipliers (ADMM) [8, 14].
  - Accelerated variants of AMA and ADMM [6, 12].
  - Preconditioned ADMM, Linearized ADMM and inexact Uzawa algorithms [2, 17].
- **Second-order decomposition methods:**
  - Dual (quasi) Newton methods [22].
  - Smoothing decomposition methods via barriers functions [15, 20, 23].

Performance of optimization algorithms

A performance metric: Time-to-reach ϵ

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

Finding the fastest algorithm within the zoo is tricky!

- heuristics & tuning parameters
- non-optimal rates & strict assumptions
- lack of precise characterizations

The optimal solution set

Optimality condition

The **optimality condition** of $\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$ (e.g., simplified (1)):

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

(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.

Finding an optimal solution

A plausible algorithmic strategy for $\min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$:

A natural minimax formulation:

$$(\mathbf{x}^\star, y^\star) \in \arg\min_{x} \max_{y\in\mathcal{Y}} \{\mathcal{L}(x,y) := f(x) + \langle y, Ax - b\rangle\}.$$

Dual subproblem: $y^*(x) \in \arg\max_{y\in\mathcal{Y}} \mathcal{L}(x,y)$

Primal problem: $x^\star \in \arg\min_x \{\mathcal{L}(y^*(x), x)\}$

A basic strategy $\Rightarrow$ Find $x^\star$ by using $y^*(x)$

- We will discuss two approaches in the sequel

Primal, Dual and Lagrangian

Using the max-form of the indicator function, primal problem can be written as

$$F^\star := \min_x \max_y \left\{\mathcal{L}(x,y) := f(x) + \langle Ax - b, y\rangle\right\}.$$

Dual problem is

$$D^\star := \max_y \min_x \left\{\mathcal{L}(x,y) := f(x) + \langle Ax - b, y\rangle\right\}.$$

$$\begin{aligned} D^\star = \max_y \min_x \{f(x) + \langle Ax - b, y\rangle\} &\leq \min_x \max_y \{f(x) + \langle Ax - b, y\rangle\} \\ &= \begin{cases} \min_x f(x) & \text{if } Ax = b, \\ +\infty & \text{otherwise} \end{cases} \end{aligned} \tag{3}$$

Here, the inequality is due to **the max-min theorem** [18].

Saddle point

Definition (Saddle point)

A point $(x^\star, y^\star) \in \mathcal{X} \times \mathbb{R}^n$ is called a saddle point of the Lagrange function $\mathcal{L}$ if

$$\mathcal{L}(x^\star, y) \leq \mathcal{L}(x^\star, y^\star) \leq \mathcal{L}(x, y^\star), \ \forall x \in \mathcal{X}, \ y \in \mathbb{R}^n.$$

Recall the minimax form:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{\mathcal{L}(x, y) := f(x) + \langle y, Ax - b\rangle\}.$$

Saddle point

Definition (Saddle point)

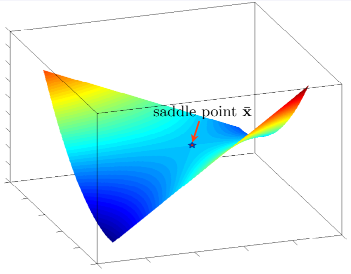
A point $(x^\star, y^\star) \in \mathcal{X} \times \mathbb{R}^n$ is called a saddle point of the Lagrange function $\mathcal{L}$ if

$$\mathcal{L}(x^\star, y) \leq \mathcal{L}(x^\star, y^\star) \leq \mathcal{L}(x, y^\star), \ \forall x \in \mathcal{X}, \ y \in \mathbb{R}^n.$$

Recall the minimax form:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \{\mathcal{L}(x, y) := f(x) + \langle y, Ax - b \rangle\}.$$

Illustration of saddle point: $\mathcal{L}(x, y) := (1/2)x^2 + y(x-1)$ in $\mathbb{R}^2$



⋆Slater's qualification condition

Slater's qualification condition

Recall $\mathrm{relint}(\mathcal{X})$ the relative interior of the **feasible set** $\mathcal{X}$. The Slater condition requires

$$\boxed{\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \;:\; \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{4}$$

⋆Slater's qualification condition

Slater's qualification condition

Recall $\mathrm{relint}(\mathcal{X})$ the relative interior of the **feasible set** $\mathcal{X}$. The Slater condition requires

$$\boxed{\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{4}$$

Special cases

- If $\mathcal{X}$ is absent, then (4) $\Leftrightarrow$ $\boxed{\exists \bar{\mathbf{x}} \ : \ \mathbf{A}\bar{\mathbf{x}} = \mathbf{b}}$.
- If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is absent, then (4) $\Leftrightarrow$ $\boxed{\mathrm{relint}(\mathcal{X}) \neq \emptyset}$.
- If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is absent and $\mathcal{X} := \{\mathbf{x} : h(\mathbf{x}) \leq 0\}$, where h is $\mathbb{R}^p \to R^q$ is convex, then

$$(4) \Leftrightarrow \boxed{\exists \bar{\mathbf{x}} \ : \ h(\bar{\mathbf{x}}) < 0.}$$

A composite reformulation

- Focus the following template in the sequel:

$$\boxed{\min_x \{f(x) : Ax = b, x \in \mathcal{X}\}}$$

- Fundamentally the same as the composite form: $\min_{x\in\mathcal{X}} f(x) + g(Ax)$

Lasso	$\mathcal{X} = \mathbb{R}^p$	$f(x) = \lambda\|x\|_1$	$g(z) = \frac{1}{n}\|z - b\|_2^2$
Square-root Lasso	$\mathcal{X} = \mathbb{R}^p$	$f(x) = \lambda\|x\|_1$	$g(z) = \frac{1}{\sqrt{n}}\|z - b\|_2$
SDP	$\mathcal{X} = \{x \succeq 0, x' = x\}$	$f(x) = \mathrm{tr}(bx)$	$g(z) = \begin{cases} 0 & \text{if } z = b \\ +\infty & \text{otherwise} \end{cases}$

Lasso is essentially "easy"

$$\boxed{\min_{x \in \mathcal{X}} f(x) + g(Ax)}$$

- Revelation: Lasso can be solved as if the problem is fully smooth!
  - **not with subgradient descent!**

- Structures in the composite form
  - g has Lipschitz gradient in ℓ_2-norm (i.e., $\|\nabla g(u) - \nabla g(v)\|_2 \leq L\|u - v\|_2$)

    Lasso: $g(x) = \frac{1}{2}\|x\|_2^2 \Rightarrow L = 1.$

  - $f : \mathcal{X} \to \mathbb{R} \cup \{+\infty\}$ has a "tractable" proximal operator

$$\mathrm{prox}_f(x) := \arg\min_{u \in \mathcal{X}} f(u) + \frac{1}{2}\|u - x\|_2^2$$

    Lasso: $f(x) = \|x\|_1, \mathcal{X} = \mathbb{R}^p \Rightarrow \mathrm{prox}_f$ is soft thresholding.

Famous Algorithms I

$$\boxed{\min_{x\in\mathcal{X}} f(x) + g(Ax)}$$

- FISTA (aka. accelerated proximal gradient method, aka. Nesterov acceleration):

  At iteration k:

$$x^{k+1} = \mathrm{prox}_{f/L\|A\|^2}\left(y^k - \frac{1}{L\|A\|^2}A^\top\nabla g(Ay^k)\right)$$
$$y^{k+1} = x^{k+1} + \frac{k+1}{k+3}\left(x^{k+1} - x^k\right)$$

- Convergence: We have

$$f(x^k) + g(Ax^k) - (f(x^\star) + g(Ax^\star)) \leq \frac{4L\|A\|^2\|x^\star - x^0\|_2^2}{(k+1)^2}$$

Conjugation of functions

Definition

Let $\mathcal{Q}$ be a predefined Euclidean space and $\mathcal{Q}^*$ be its dual space. Given a proper, closed and convex function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, the function $f^* : \mathcal{Q}^* \to \mathbb{R} \cup \{+\infty\}$ such that

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} \left\{ \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \right\}$$

is called the Fenchel conjugate (or conjugate) of f.

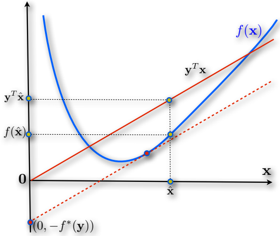


Figure: The conjugate function $f^*(\mathbf{y})$ is the maximum gap between the linear function $\mathbf{x}^T\mathbf{y}$ (red line) and $f(\mathbf{x})$.

- f^* is a convex and lower, semicontinuous function by construction (as the supremum of affine functions of $\mathbf{y}$).
- The conjugate of the conjugate of a convex function f is ... the same function f; i.e., $f^{**} = f$ for $f \in \mathcal{F}(\mathcal{Q})$.

A useful minimax reformulation for the general case

$$\boxed{\min_{x\in\mathcal{X}} f(x) + g(Ax)}$$

- If $0 \in \mathrm{ri}(\mathrm{dom}g - A\mathrm{dom}f)$ then the optimization problem is equivalent to

$$\max_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} f(x) + \langle y, Ax\rangle - g^*(y)$$

  where g^* is the Fenchel conjugate of g: $g^*(y) := \max_x \langle x, y\rangle - g(x)$.

  - Constrained case: $g(z) = \begin{cases} 0 & \text{if } z = c \\ +\infty & \text{otherwise} \end{cases}$, and hence, $g^*(y) = \langle c, y\rangle$

Duality gap

- The duality gap:

$$G(x,y) = f(x) + g(Ax) + g^*(y) + f^*(-A^\top y)$$

$$= \max_{\bar{y}\in\mathcal{Y}} \Big(f(x) + \langle \bar{y}, Ax\rangle - g^*(\bar{y}) \Big) - \min_{\bar{x}\in\mathcal{X}} \Big(-g^*(y) + \langle \bar{x}, A^\top y\rangle + f(\bar{x}) \Big)$$

  - Note the symmetric roles between (f, g, A) and $(-g^*, -f^*, A^\top)$

- Useful properties:
  - Convex as a function of (x, y)
  - $G(x,y) = 0$ iff $(x,y) = (x^\star, y^\star)$

⋆Famous algorithms II

- Chambolle-Pock method (dual perspective):

  At iteration k:

$$x^{k+1} = \arg\min_{x\in\mathcal{X}} f(x) + \langle y^k, Ax - c\rangle + \frac{\beta}{2}\left\|x - x^k\right\|^2$$
$$y^{k+1} = y^k + \frac{\beta - \epsilon}{\|A\|^2_{\mathcal{X},\mathcal{Y}}}\left(A(2x^{k+1} - x^k) - c\right)$$

- Convergence: We have

$$G(x^k, y^k) \leq \frac{1}{k}\left(\frac{\beta}{2}D^2_{\mathcal{X}} + \frac{\|A\|^2}{2(\beta-\epsilon)}D^2_{\mathcal{Y}}\right)$$

  where $D_{\mathcal{X}}$ is the diameter of dom$f$ and $D_{\mathcal{Y}}$ is the diameter of domg^*.

A Primer on Smoothing

- Assuming that g admits max-form

$$g(z) = \max_{y \in \mathcal{Y}} \left(\langle z, y \rangle - g^*(y) \right). \tag{5}$$

- A smoothed estimate of g by Nesterov around a center point $\dot{y}$:

$$g_{\beta}(z; \dot{y}) = \max_{y \in \mathcal{Y}} \left(\langle z, y \rangle - g^*(y) - \frac{\beta}{2} \|y - \dot{y}\|^2 \right)$$

- The approximation guarantee

$$g_{\beta}(z; \dot{y}) \leq g(z) \leq g_{\beta}(z; \dot{y}) + \frac{\beta}{2} D^2, \tag{6}$$

where $D = \max_{y \in \mathrm{dom}(g^*)} \|y - \dot{y}\|$.

Examples

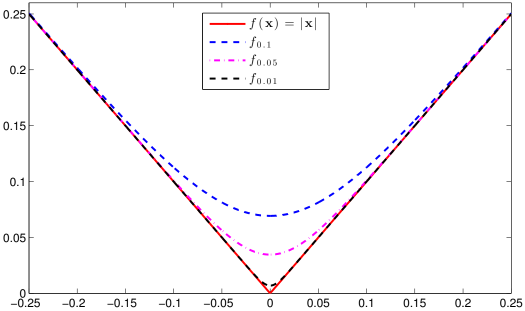
- Absolute value function in max-form

$$g(x) = |x| = \max_{-1 \leq y \leq 1} xy.$$

- Let $\dot{y} = 0$,

$$g_\beta(x) = \max_{-1 \leq y \leq 1} \left(xy - \frac{\beta}{2} y^2 \right) = \begin{cases} \frac{x^2}{2\beta}, & |x| \leq \beta \\ |x| - \frac{\beta}{2}, & |x| > \beta \end{cases}.$$

- Smoothed ℓ_1-norm is the so-called Huber loss.



Examples

- Constrained case $i.e.$ when g is an indicator function:

$$g(z) = \delta_{\{c\}}(z) = \begin{cases} 0 & \text{if } z = c \\ +\infty & \text{otherwise} \end{cases}, \text{ and hence, } g^*(y) = \langle c, y \rangle$$

- g_β is differentiable wrt z and $\nabla_z g_\beta$ is $\frac{1}{\beta}$-Lipschitz

- $g_\beta(Ax^k, \dot{y}) = \langle \dot{y}, Ax^k - c \rangle + \frac{1}{2\beta} \left\| Ax^k - c \right\|^2$

Efficiency considerations from the dual problem

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

Subgradient method for the nonsmooth problem

Assume that the following conditions

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

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

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

Efficiency considerations from the dual problem

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

Subgradient method for the nonsmooth problem

Assume that the following conditions

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

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

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

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

Efficiency considerations from the dual problem

Gradient method
1. Choose $y^0 \in \mathbb{R}^n$.
2. For $k = 0, 1, \cdots$, perform:
$y^{k+1} = y^k + \frac{1}{L}\nabla d(y^k)$,
where L is the Lipschitz constant.

Subgradient method for the nonsmooth problem

Assume that the following conditions

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

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

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

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) $d(y)$ has Lipschitz continuous gradient iff

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

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

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

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

Efficiency considerations from the dual problem

Gradient method

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

Subgradient method for the nonsmooth problem

Assume that the following conditions

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

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

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

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) $d(y)$ has Lipschitz continuous gradient iff

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

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

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

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

This is NOT the best we can do.

There exists a complexity lower-bound

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

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

Efficiency considerations from the dual problem

Accelerated gradient method

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

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

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

Subgradient method for the nonsmooth problem

Assume that the following conditions

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

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

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

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) $d(y)$ has Lipschitz continuous gradient iff

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

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

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

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

This is NEARLY the best we can do.

There exists a complexity lower-bound

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

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

Number of iterations: From $\mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ to $\mathcal{O}\left(\frac{1}{\epsilon}\right)$

When can the function have Lipschitz gradient?

When $g^*(y)$ is γ-strongly convex, the conjugate function $g(Ax)$ is $\frac{\|\mathbf{A}\|^2}{\gamma}$-Lipschitz gradient.

(Strong convexity) $g^*(y)$ is γ-strongly convex iff $g^*(y) - \frac{\gamma}{2}\|y\|_2^2$ is convex.

Number of iterations: From $\mathcal{O}\left(\frac{1}{\bar{\epsilon}^2}\right)$ to $\mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$

When can the function have Lipschitz gradient?

When $g^*(y)$ is γ-strongly convex, the conjugate function $g(Ax)$ is $\frac{\|\mathbf{A}\|^2}{\gamma}$-Lipschitz gradient.

(Strong convexity) $g^*(y)$ is γ-strongly convex iff $g^*(y) - \frac{\gamma}{2}\|y\|_2^2$ is convex.

A simple idea: Apply Nesterov's smoothing [16]

$$g_\gamma(Ax) = \max_y \langle Ax, y\rangle - g^*(y) - \frac{\gamma}{2}\|y\|_2^2$$

1. $\nabla g_\gamma(Ax) = A^\top y^*_\gamma(Ax)$
2. $g_\gamma(Ax) \leq g(Ax) \leq g_\gamma(Ax) + \gamma\mathcal{D}_\mathcal{Y}$, where $\mathcal{D}_\mathcal{Y} = \max_{y\in\mathcal{Y}} \frac{1}{2}\|y\|_2^2$.
3. x^k of AGM on $g_\gamma(Ax)$ has
   $g^\star - g(Ax^k) \leq \gamma\mathcal{D}_\mathcal{Y} + g^\star_\gamma - g_\gamma(Ax^k) \leq \gamma\mathcal{D}_\mathcal{Y} + \frac{2\|A\|^2R^2}{\gamma(k+2)^2}$.
4. We minimize the upperbound wrt γ and obtain $g^\star - g(Ax^k) \leq \bar{\epsilon}$ with $k = \mathcal{O}\left(\frac{1}{\bar{\epsilon}}\right)$.

Per-iteration time: The key role of the prox-operator

Smoothed function: $g_\gamma(Ax) = \max_y \langle Ax, y\rangle - g^*(y) - \frac{\gamma}{2}\|y\|_2^2$

$$y^*_\gamma(Ax) := \mathrm{prox}^{\mathcal{X}}_{g^*/\gamma}\left(-\frac{1}{\gamma}Ax\right)$$

Per-iteration time: The key role of the prox-operator

Smoothed function: $g_\gamma(Ax) = \max_y \langle Ax, y\rangle - g^*(y) - \frac{\gamma}{2}\|y\|_2^2$

$$y^*_\gamma(Ax) := \mathrm{prox}^{\mathcal{X}}_{g^*/\gamma}\left(-\frac{1}{\gamma}Ax\right)$$

Definition (Prox-operator)

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

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 $f(\mathbf{z}) = \|\mathbf{z}\|_1$, then the prox-operator performs coordinate-wise soft-thresholding by 1.

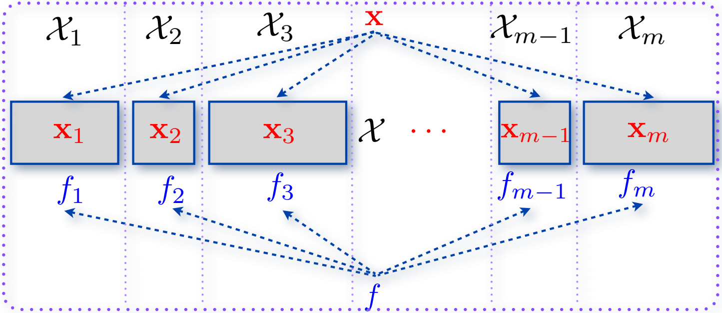
Decomposability

Decomposable structure

The function f and the feasible set $\mathcal{X}$ have the following 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, $\mathbf{x}_i$ is a sub-vector (component) of $\mathbf{x}$, $f_i : \mathbb{R}^{p_i} \to \mathbb{R} \cup \{+\infty\}$ is convex and $\sum_{i=1}^{m} p_i = p$.



A first attempt

- Nesterov's smooth minimization of non-smooth functions approach:

  Choose $\beta > 0$ and $\dot{y}$.

  Run FISTA on $x \mapsto f(x) + g_\beta(Ax, \dot{y})$ as a proxy for $f(x) + g(Ax)$.

- Convergence:

$$f(x^k) + g_\beta(Ax^k, \dot{y}) - \big(f(x^\star) + g_\beta(Ax^\star)\big) \leq \frac{4\|A\|^2 \left\|x^0 - x^\star\right\|^2}{\beta(k+1)^2}$$

$$f(x^k) + g(Ax^k) - (f(x^\star) + g(Ax^\star)) \leq \frac{4\|A\|^2 \left\|x^0 - x^\star\right\|^2}{\beta(k+1)^2} + \beta D_{\mathcal{Y}}$$

Our fundamental theorem

- Recall the duality gap:

$$G(x,y) = f(x) + g(Ax) + g^*(y) + f^*(-A^\top y)$$

$$= \max_{\bar{y}\in\mathcal{Y}} \Big(f(x) + \langle \bar{y}, Ax\rangle - g^*(\bar{y})\Big) - \min_{\bar{x}\in\mathcal{X}} \Big(- g^*(y) + \langle \bar{x}, A^\top y\rangle + f(\bar{x})\Big)$$

- Denote the (primal) smoothed gap function at $y^\star$ as

$$S_\beta(x,\dot{y}) := f(x) + g_\beta(Ax;\dot{y}) - f(x^\star)$$

Theorem

If β and $S_\beta(x,\dot{y})$ are small, we have an approximate solution:

$$\|Ax - c\| \leq \beta\Big[\|y^\star - \dot{y}\| + \Big(\|y^\star - \dot{y}\|^2 + 2\beta^{-1}S_\beta(x;\dot{y})\Big)^{1/2}\Big]$$

$$f(x) - f(x^\star) \geq -\|y^\star\|\|Ax - c\|$$

$$f(x) - f(x^\star) \leq S_\beta(x,\dot{y}) + \|y^\star\|\|Ax - c\| + \frac{\beta}{2}\|y^\star - \dot{y}\|^2$$

Accelerated Smoothed GAp ReDuction algorithm (ASGARD)

Idea: FISTA on $f(x) + g_\beta(Ax;\dot{y})$ and continuation on β

For $k = 0$ **to** $k_{\max}$:

$$y^*_{\beta_{k+1}}(A\hat{x}^k;\dot{y}) = \arg\max_{y\in\mathcal{Y}} \langle A\hat{x}^k, y\rangle - g^*(\hat{y}) - \frac{\beta_{k+1}}{2}\|y-\dot{y}\|^2$$

$$\bar{x}^{k+1} = \mathrm{prox}_{\beta_{k+1}\|A\|^{-2}f}\left(\hat{x}^k - \beta_{k+1}\|A\|^{-2}A^\top y^*_{\beta_{k+1}}(A\hat{x}^k;\dot{y})\right)$$

$$\hat{x}^{k+1} = \bar{x}^{k+1} + \tfrac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\bar{x}^{k+1} - \bar{x}^k)$$

$\tau_{k+1} \in (0,1)$ root of $\tau^3 + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$

$$\beta_{k+2} = \tfrac{\beta_{k+1}}{1+\tau_{k+1}}$$

End for

Convergence theorem

Theorem

The iterates of ASGARD drive the smoothed gap to zero: $S_{\beta_k}(\bar{x}^k, \dot{y}) = \mathcal{O}(1/k)$, and also provides a $\mathcal{O}(1/k)$ convergence guarantee in function value as well as feasibility:

$$\left\|A\bar{x}^k - c\right\| \leq \frac{\beta_1}{k+1}\left[\left\|y^\star - \dot{y}\right\| + \sqrt{\|y^\star - \dot{y}\|^2 + \frac{\|A\|^2}{\beta_1^2}\|\bar{x}^0 - x^\star\|^2}\right]$$

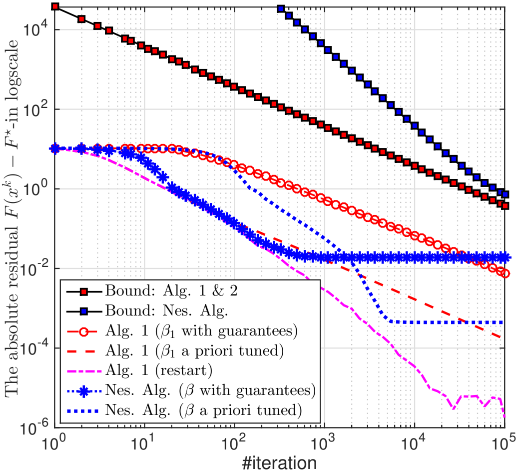
$$f(\bar{x}^k) - f(x^\star) \geq -\|y^\star\|\|Ax - c\|$$

$$f(\bar{x}^k) - f(x^\star) \leq \frac{1}{k}\frac{\|A\|^2}{2\beta_1}\left\|\bar{x}^0 - x^\star\right\|^2 + \left\|y^\star\right\|\left\|A\bar{x}^k - c\right\| + \frac{\beta_1}{k+1}\left\|y^\star - \dot{y}\right\|^2$$

Square-root Lasso: A comparison with Nesterov's smoothing

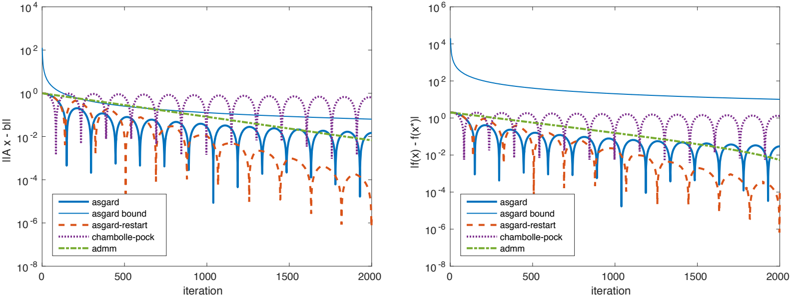
$$\min_x F(x) := \frac{1}{\sqrt{m}} \|Ax - b\|_2 + \lambda \|x\|_1$$

Tune β using $\|x^\star\|_2 \leq \frac{\|b\|_2}{\lambda \sqrt{m}}$



A degenerate LP problem: Guarantees matter

$$
\begin{aligned}
\min_{x\in\mathbb{R}^n} \quad & 2x_n \\
\text{s.t.} \quad & x_n \geq 0, \qquad \textstyle\sum_{k=1}^{n-1} x_k = 1 \\
& x_n - \textstyle\sum_{k=1}^{n-1} x_k = 0 \quad (2 \leq j \leq d) \\
(n = 10, \quad & d = 200)
\end{aligned}
$$



A versatile framework

- Extensions
  - Restart
  - Augmented Lagrangian smoother
  - ADSGARD: The dual perspective
  - Linearization of smooth parts of the objective
  - Line search
  - Use of old gradients (aka gradient sliding)
  - Splitting the smoothed gap (ADMM-version)
  - Random coordinate descent updates with continuation

Accelerated Augmented Lagragian method

Idea #1: Smooth the dual (i.e., f^*) with $\|x\|_{\mathcal{X}} = \|Ax\|_{\mathcal{Y},*}$ and $\dot{x} = x^\star$.

Idea #2: FISTA on $-g^*(y) - f^*_\beta(-A^\top y)$ and continuation on β

Algorithm ASALGARD: FISTA in disguise for the dual & averaging for the primal

$$
\begin{aligned}
\hat{y}^k &= (1-\tau_k)\bar{y}^k + \tau_k y^*_{\beta_k}(A\bar{x}^k; \dot{y})\\
\hat{x}^*_{\gamma_0}(\hat{y}^k) &= \arg\min_{x\in\mathcal{X}} \left\{ f(x) + \langle \hat{y}^k, Ax - c\rangle + \frac{\gamma_0}{2}\|Ax - c\|^2_{\mathcal{Y},*} \right\}\\
\bar{y}^{k+1} &= \hat{y}^k + \gamma_0(A\hat{x}^*_{\gamma_0}(\hat{y}^k) - c)\\
\bar{x}^{k+1} &= (1-\tau_k)\bar{x}^k + \tau_k \hat{x}^*_{\gamma_0}(\hat{y}^k)\\
\tau_{k+1} &\in (0,1) \text{ root of } \tau^2 + \tau_k^2\tau - \tau_k^2 = 0\\
\beta_{k+2} &= (1-\tau_{k+1})\beta_{k+1}
\end{aligned}
$$

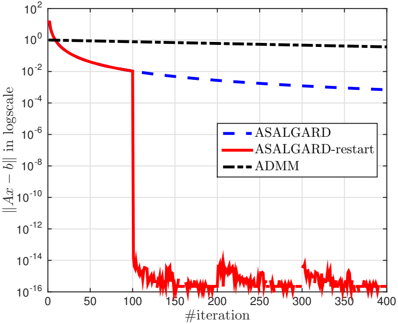
Convergence: We have

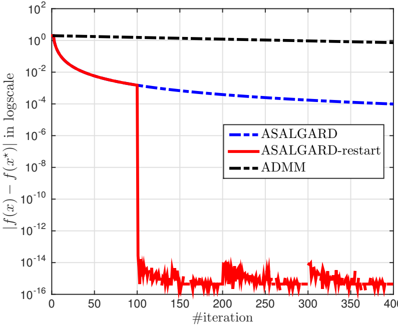
$$
-\frac{8\|y^\star\|_{\mathcal{Y}}\|y^\star - \dot{y}\|_{\mathcal{Y}}}{\gamma_0(k+2)^2} \le f(\bar{x}^k) - f^\star \le \frac{8\|y^\star\|_{\mathcal{Y}}\|y^\star - \dot{y}\|_{\mathcal{Y}} + 2\|y^\star - \dot{y}\|^2}{\gamma_0(k+2)^2}
$$

$$
\|A\bar{x}^k - c\|_{\mathcal{Y},*} \le \frac{8\|y^\star - \dot{y}\|_{\mathcal{Y}}}{\gamma_0(k+2)^2}
$$

ASALGARD on the same degenerate LP problem

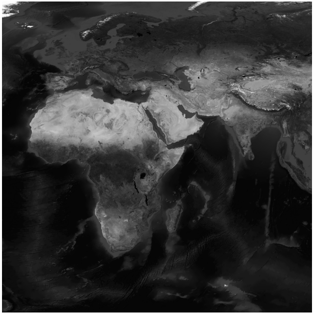
$$
\begin{aligned}
\min_{x\in\mathbb{R}^n} \quad & 2x_n \\
\text{s.t.} \quad & x_n \geq 0, \qquad \textstyle\sum_{k=1}^{n-1} x_k = 1 \\
& x_n - \textstyle\sum_{k=1}^{n-1} x_k = 0 \quad (2 \leq j \leq d) \\
(n = 10, \quad & d = 200)
\end{aligned}
$$



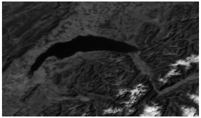


Tree sparsity example: 1:100-compressive sensing

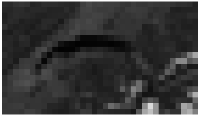
World [1Gpix]



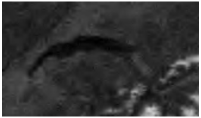
Lac Léman



World [10Mpix]

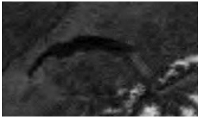


wavelet sparse



PNSR $= 31.83$db

wavelet tree-sparse



PNSR $= 32.48$db

ASALGARD:

Iterations: 113
PD gap: 1e-8
Applications of $(\mathbf{A}, \mathbf{A}^T)$: $(684, 570)$

$$\min_{x \in \mathbb{R}^p} \quad f(x) := \sum_{\mathcal{G}_i \in \mathfrak{G}} \|x_{\mathcal{G}_i}\|_\infty$$
$$\text{s.t.} \quad Ax = c.$$

References I

[1] H.H. Bauschke and P. Combettes.
Convex analysis and monotone operators theory in Hilbert spaces.
Springer-Verlag, 2011.

[2] A. Chambolle and T. Pock.
A first-order primal-dual algorithm for convex problems with applications to imaging.
Journal of Mathematical Imaging and Vision, 40(1):120–145, 2011.

[3] G. Chen and M. Teboulle.
A proximal-based decomposition method for convex minimization problems.
Math. Program., 64:81–101, 1994.

[4] P. L. Combettes and V. R. Wajs.
Signal recovery by proximal forward-backward splitting.
Multiscale Model. Simul., 4:1168–1200, 2005.

[5] D. Davis.
Convergence rate analysis of the forward-Douglas-Rachford splitting scheme.
UCLA CAM report 14-73, 2014.

References II

[6] D. Davis and W. Yin.
Faster convergence rates of relaxed Peaceman-Rachford and ADMM under regularity assumptions.
UCLA CAM report 14-58, 2014.

[7] D. Davis and W. Yin.
A three-operator splitting scheme and its optimization applications.
Tech. Report., 2015.

[8] J. Eckstein and D. Bertsekas.
On the Douglas - Rachford splitting method and the proximal point algorithm for maximal monotone operators.
Math. Program., 55:293–318, 1992.

[9] J. E. Esser.
Primal-dual algorithm for convex models and applications to image restoration, registration and nonlocal inpainting.
Phd. thesis, University of California, Los Angeles, Los Angeles, USA, 2010.

[10] D. Gabay and B. Mercier.
A dual algorithm for the solution of nonlinear variational problems via finite element approximation.
Computers & Mathematics with Applications, 2(1):17 – 40, 1976.

References III

[11] T. Goldstein, E. Esser, and R. Baraniuk.
Adaptive Primal-Dual Hybrid Gradient Methods for Saddle Point Problems.
Tech. Report., http://arxiv.org/pdf/1305.0546v1.pdf:1–26, 2013.

[12] T. Goldstein, B. ODonoghue, and S. Setzer.
Fast Alternating Direction Optimization Methods.
SIAM J. Imaging Sci., 7(3):1588–1623, 2012.

[13] B. He and X. Yuan.
Convergence analysis of primal-dual algorithms for saddle-point problem: from contraction perspective.
SIAM J. Imaging Sciences, 5:119–149, 2012.

[14] B.S. He and X.M. Yuan.
On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method.
SIAM J. Numer. Anal., 50:700–709, 2012.

[15] I. Necoara and J.A.K. Suykens.
Interior-point lagrangian decomposition method for separable convex optimization.
J. Optim. Theory and Appl., 143(3):567–588, 2009.

References IV

[16] Y. Nesterov.
Smooth minimization of non-smooth functions.
Math. Program., 103(1):127–152, 2005.

[17] Y. Ouyang, Y. Chen, G. LanG. Lan., and E. JR. Pasiliao.
An accelerated linearized alternating direction method of multiplier.
Tech, 2014.

[18] R. T. Rockafellar.
Convex Analysis, volume 28 of *Princeton Mathematics Series*.
Princeton University Press, 1970.

[19] Q. Tran-Dinh and V. Cevher.
Constrained convex minimization via model-based excessive gap.
In *Proc. the Neural Information Processing Systems Foundation conference (NIPS2014)*, pages 1–9, Montreal, Canada, December 2014.

[20] Q. Tran-Dinh, I. Necoara, C. Savorgnan, and M. Diehl.
An Inexact Perturbed Path-Following Method for Lagrangian Decomposition in Large-Scale Separable Convex Optimization.
SIAM J. Optim., 23(1):95–125, 2013.

References V

[21] P. Tseng.
Applications of splitting algorithm to decomposition in convex programming and variational inequalities.
SIAM J. Control Optim., 29:119–138, 1991.

[22] E. Wei, A. Ozdaglar, and A.Jadbabaie.
A Distributed Newton Method for Network Utility Maximization.
http://web.mit.edu/asuman/www/publications.htm, 2011.

[23] G. Zhao.
A Lagrangian dual method with self-concordant barriers for multistage stochastic convex programming.
Math. Progam., 102:1–24, 2005.