Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Constrained convex minimization I

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

EE-556 (Fall 2015)

lions@epfl

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

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 and Volkan 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.
- 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

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

- 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

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

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

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

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

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

- 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 $\mathbf{x}^\star \in \mathcal{X}$

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, ...*

Primal-dual methods for (1):

Plenty . . .

- Variants of the **Arrow-Hurwitz's method**:
  - Chambolle-Pock's algorithm [4], and its variants, e.g., He-Yuan's variant [17].
  - Primal-dual Hybrid Gradient (PDHG) method and its variants [13, 15].
  - Proximal-based decomposition (Chen-Teboulle's algorithm) [5].
- Splitting techniques from monotone inclusions:
  - Primal-dual splitting algorithms [3, 6, 28, 7, 8].
  - Three-operator splitting [9].
- Dual splitting techniques:
  - Alternating minimization algorithms (AMA) [14, 28].
  - Alternating direction methods of multipliers (ADMM) [11, 18].
  - Accelerated variants of AMA and ADMM [8, 16].
  - Preconditioned ADMM, Linearized ADMM and inexact Uzawa algorithms [4, 23].
- **Second-order decomposition methods:**
  - Dual (quasi) Newton methods [29].
  - Smoothing decomposition methods via barriers functions [20, 27, 30].

Performance of optimization algorithms

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (1) is **impracticable**
- Algorithms seek $\mathbf{x}_\epsilon^\star$ 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 $\mathbf{x}_\epsilon^\star$ 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 $\mathbf{x}_\epsilon^\star$ 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)!

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

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

Numerical ϵ-accuracy

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

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

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

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

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

Numerical ϵ-accuracy

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

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

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

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

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

Our definition of ϵ-accurate solutions [25]

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

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

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

Numerical ϵ-accuracy

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

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

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

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

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

Our definition of ϵ-accurate solutions [25]

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

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

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

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

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

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

In the sequel: Heuristic-free optimal *first-order* primal-dual / ADMM / AMA methods

Outline

The proximal way

Establishing correctness

Efficiency considerations

Back to the primal

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.

Example: Basis pursuit

Example (Basis pursuit)

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

Note:

- $f(\mathbf{x}) := \|\mathbf{x}\|_1$ is nonsmooth, for any $\mathbf{v} \in \partial f(\mathbf{x})$ we have $v_i = +1$ if $x_i > 0$, $v_i = -1$ if $x_i < 0$ and $v_i \in (-1, 1)$ if $x_i = 0$.
- Since $\mathcal{X} \equiv \mathbb{R}^p$, we have $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{0\}$ for all $\mathbf{x}$.

Example: Basis pursuit

Example (Basis pursuit)

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

Note:

- $f(\mathbf{x}) := \|\mathbf{x}\|_1$ is nonsmooth, for any $\mathbf{v} \in \partial f(\mathbf{x})$ we have $v_i = +1$ if $x_i > 0$, $v_i = -1$ if $x_i < 0$ and $v_i \in (-1, 1)$ if $x_i = 0$.
- Since $\mathcal{X} \equiv \mathbb{R}^p$, we have $\mathcal{N}_{\mathcal{X}}(\mathbf{x}) = \{0\}$ for all $\mathbf{x}$.

Optimality condition

The **optimality condition** of (2) becomes

$$\begin{cases} 0 \in \partial f(\mathbf{x}^\star) + \mathbf{A}^T\lambda^\star \\ 0 = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases} \quad \Leftrightarrow \quad \begin{cases} (\mathbf{A}^T\lambda^\star)_i = -1 & \text{if } x_i^\star > 0,\ 1 \le i \le p \\ (\mathbf{A}^T\lambda^\star)_i = +1 & \text{if } x_i^\star < 0,\ 1 \le i \le p \\ (\mathbf{A}^T\lambda^\star)_i \in (-1, 1) & \text{if } x_i^\star = 0,\ 1 \le i \le p \\ \mathbf{A}\mathbf{x}^\star = \mathbf{b}. \end{cases}$$

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:

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

Lagrangian subproblem: $\mathbf{x}^*(\lambda)\in \arg\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}(\mathbf{x},\lambda)$

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

- λ is called the Lagrange multiplier.
- The function $d(\lambda)$ is called the dual function, and it is concave!
- The optimal dual objective value is $d^\star=d(\lambda^\star)$.

A basic strategy $\Rightarrow$ Find $\lambda^\star$ and then solve for $\mathbf{x}^\star=\mathbf{x}^*(\lambda^\star)$

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:

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

Lagrangian subproblem: $\mathbf{x}^*(\lambda)\in \arg\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}(\mathbf{x},\lambda)$

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

- λ is called the Lagrange multiplier.
- The function $d(\lambda)$ is called the dual function, and it is concave!
- The optimal dual objective value is $d^\star=d(\lambda^\star)$.

A basic strategy $\Rightarrow$ Find $\lambda^\star$ and then solve for $\mathbf{x}^\star=\mathbf{x}^*(\lambda^\star)$

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 $\lambda^\star_{\bar{\epsilon}}\rightarrow\mathbf{x}^\star_\epsilon$ (i.e., $\bar{\epsilon}(\epsilon)$), where ϵ is for the original constrained problem (1)

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:

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

Lagrangian subproblem: $\mathbf{x}^*(\lambda)\in\arg\min_{\mathbf{x}\in\mathcal{X}}\mathcal{L}(\mathbf{x},\lambda)$

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

- λ is called the Lagrange multiplier.
- The function $d(\lambda)$ is called the dual function, and it is concave!
- The optimal dual objective value is $d^\star=d(\lambda^\star)$.

A basic strategy $\Rightarrow$ Find $\lambda^\star$ and then solve for $\mathbf{x}^\star=\mathbf{x}^*(\lambda^\star)$

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)

Outline

The proximal way

Establishing correctness

Efficiency considerations

Back to the primal

Back to the the minimax formulation

The dual function and the dual problem revisited

- **Dual function:**

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\}. \tag{3}$$

  Let $\mathbf{x}^*(\lambda)$ be a solution of (3) then $d(\lambda)$ is finite if $x^*(\lambda)$ exists. $d(\cdot)$ is concave and possibly nonsmooth.

- **Dual problem**: The following dual problem is convex

$$\boxed{d^\star := \max_{\mathbf{x}\in\mathbb{R}^n} d(\lambda)} \tag{4}$$

Back to the the minimax formulation

The dual function and the dual problem revisited

- **Dual function:**

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \{\mathcal{L}(\mathbf{x},\lambda) := f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\}. \tag{3}$$

  Let $\mathbf{x}^*(\lambda)$ be a solution of (3) then $d(\lambda)$ is finite if $x^*(\lambda)$ exists. $d(\cdot)$ is concave and possibly nonsmooth.

- **Dual problem**: The following dual problem is convex

$$\boxed{d^\star := \max_{\mathbf{x}\in\mathbb{R}^n} d(\lambda)} \tag{4}$$

The minimax formulation

$$\begin{aligned} d^\star &= \max_{\lambda\in\mathbb{R}^n} d(\lambda) = \max_{\lambda\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathcal{X}} \{f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\} \\ &\leq \min_{\mathbf{x}\in\mathcal{X}} \max_{\lambda\in\mathbb{R}^n} \{f(\mathbf{x}) + \lambda^T(\mathbf{A}\mathbf{x}-\mathbf{b})\} = \begin{cases} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) & \text{if } \mathbf{A}\mathbf{x}=\mathbf{b}, \\ +\infty & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Here, the inequality is due to **the max-min theorem** [24].

Example: Strictly convex quadratic programming

Strictly convex quadratic programming

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & (1/2)\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{h}^T\mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array}$$

where $\mathbf{H}$ is symmetric positive definite.

Example: Strictly convex quadratic programming

Strictly convex quadratic programming

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & (1/2)\mathbf{x}^T\mathbf{H}\mathbf{x}+\mathbf{h}^T\mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x}=\mathbf{b}. \end{array}$$

where $\mathbf{H}$ is symmetric positive definite.

Dual problem is also a strictly convex quadratic program

- Lagrange function $\mathcal{L}(\mathbf{x},\lambda) := (1/2)\mathbf{x}^T\mathbf{H}\mathbf{x}+(\mathbf{A}^T\lambda+\mathbf{h})^T\mathbf{x}-\mathbf{b}^T\lambda.$
- Dual function:

$$d(\lambda)=\min_{\mathbf{x}\in\mathbb{R}^p}\{(1/2)\mathbf{x}^T\mathbf{H}\mathbf{x}+(\mathbf{A}^T\lambda+\mathbf{h})^T\mathbf{x}-\mathbf{b}^T\lambda\}$$

- Since $\mathbf{x}^\star(\lambda)=-\mathbf{H}^{-1}(\mathbf{A}^T\lambda+\mathbf{h})$, we can obtain $d(\lambda)$ explicitly as

$$d(\lambda)=-(1/2)\lambda^T(\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)\lambda-(\mathbf{b}+\mathbf{A}\mathbf{H}^{-1}\mathbf{h})^T\lambda.$$

- Dual problem (unconstrained):

$$d^\star := \max_{\lambda\in\mathbb{R}^n} d(\lambda) \quad\Leftrightarrow\quad \min_{\lambda\in\mathbb{R}^n} \frac{1}{2}\lambda^T(\mathbf{A}\mathbf{H}^{-1}\mathbf{A}^T)\lambda+(\mathbf{b}+\mathbf{A}\mathbf{H}^{-1}\mathbf{h})^T\lambda.$$

Example: Nonsmoothness of the dual function

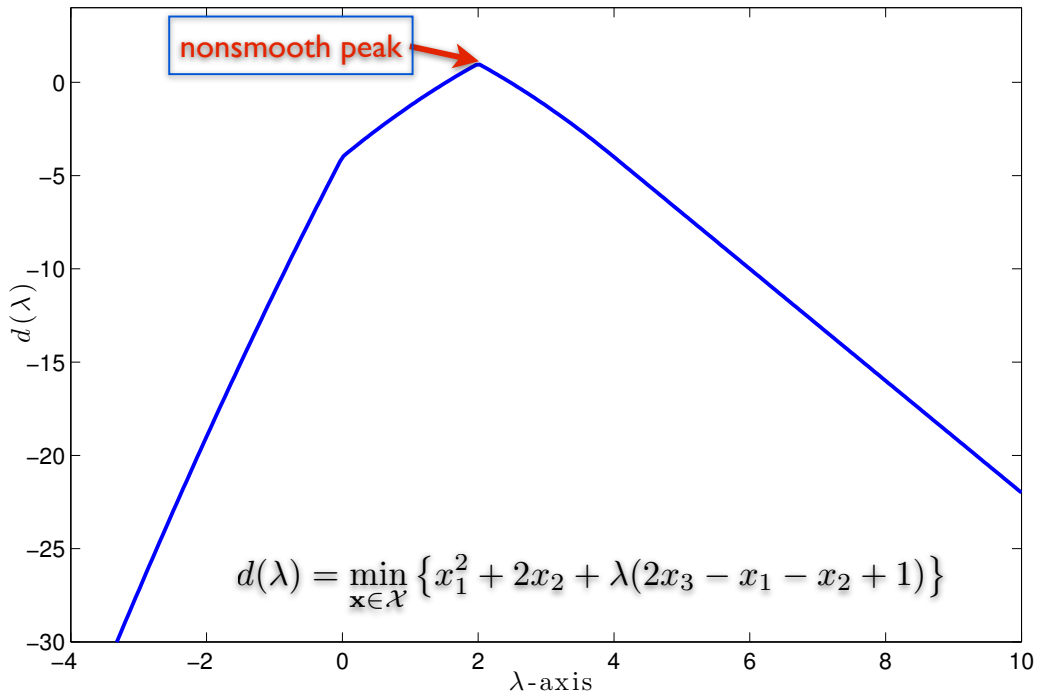
Consider a constrained convex problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^3} \quad & \{f(\mathbf{x}) := x_1^2 + 2x_2\}, \\ \text{s.t.} \quad & 2x_3 - x_1 - x_2 = 1, \\ & \mathbf{x} \in \mathcal{X} := [-2,2] \times [-2,2] \times [0,2]. \end{aligned}$$

The **dual function** is defined as

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \{x_1^2 + 2x_2 + \lambda(2x_3 - x_1 - x_2 - 1)\}$$

is concave and nonsmooth as illustrated in the figure below.



Saddle point

Definition (Saddle point)

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

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

Recall the minimax form:

$$\max_{\lambda} \min_{\mathbf{x} \in \mathcal{X}} \{\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T (\mathbf{A}\mathbf{x} - \mathbf{b})\}. \qquad ((3))$$

Saddle point

Definition (Saddle point)

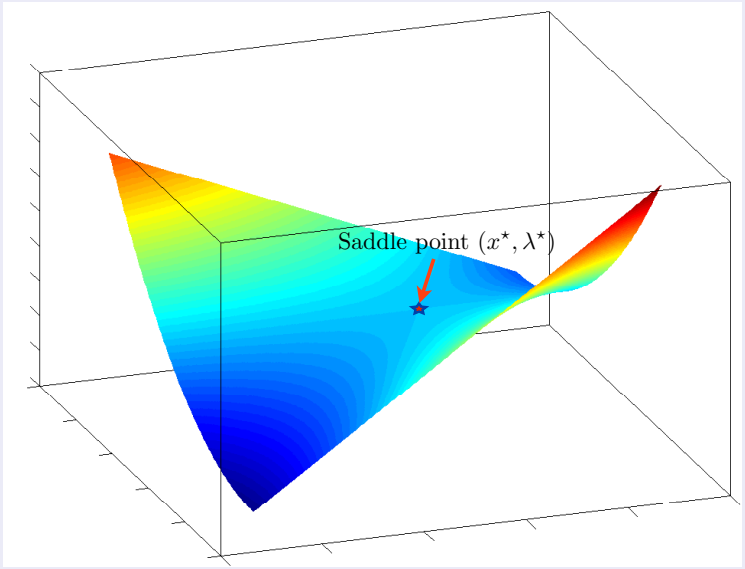
A point $(\mathbf{x}^\star, \lambda^\star) \in \mathcal{X} \times \mathbb{R}^n$ is called a saddle point of the Lagrange function $\mathcal{L}$ if

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

Recall the minimax form:

$$\max_{\lambda} \min_{\mathbf{x} \in \mathcal{X}} \{\mathcal{L}(\mathbf{x}, \lambda) := f(\mathbf{x}) + \lambda^T (\mathbf{A}\mathbf{x} - \mathbf{b})\}. \qquad ((3))$$

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

$$\boxed{\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{6}$$

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

$$\boxed{\mathrm{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{6}$$

Special cases

- If $\mathcal{X}$ is absent, then (6) $\Leftrightarrow$ $\boxed{\exists \bar{\mathbf{x}} \ : \ \mathbf{A}\bar{\mathbf{x}} = \mathbf{b}}$.
- If $\mathbf{A}\mathbf{x} = \mathbf{b}$ is absent, then (6) $\Leftrightarrow$ $\boxed{\mathrm{relint}(\mathcal{X}) \neq \emptyset}$.
- 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

$$(6) \Leftrightarrow \boxed{\exists \bar{\mathbf{x}} \ : \ h(\bar{\mathbf{x}}) < 0.}$$

Example: Slater's condition

Example

Let us consider the feasible set $\mathcal{D}_\alpha := \mathcal{X} \cap \mathcal{A}_\alpha$ as

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 \leq 1\} \ \mathcal{A}_\alpha := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1 + x_2 = \alpha\},$$

where $\alpha \in \mathbb{R}$.

Example: Slater's condition

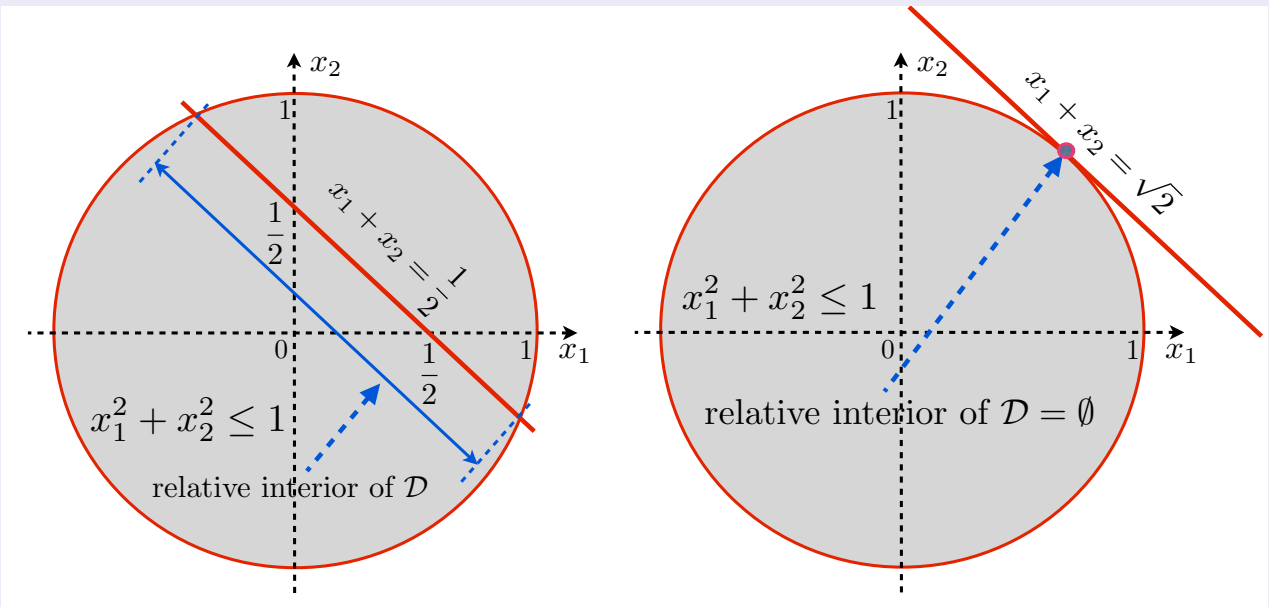
Example

Let us consider the feasible set $\mathcal{D}_\alpha := \mathcal{X} \cap \mathcal{A}_\alpha$ as

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 + x_2^2 \leq 1\} \ \mathcal{A}_\alpha := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1 + x_2 = \alpha\},$$

where $\alpha \in \mathbb{R}$.

Slater's condition holds and does not hold



$\mathcal{D}_{1/2}$ satisfies Slater's condition – $\mathcal{D}_{\sqrt{2}}$-does not satisfy Slater's condition

Necessary and sufficient condition

Theorem (Necessary and sufficient optimality condition)

Under Slater's condition (6)*:* $\operatorname{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset$*, the* ***KKT condition*** (2)

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

is necessary and sufficient for a point $(\mathbf{x}^\star, \lambda^\star) \in \mathcal{X} \times \mathbb{R}^n$ *being an optimal solution for the primal problem* (1) *and dual problem* (4)*:*

$$f^\star := \left\{ \begin{array}{ll} \min\limits_{\mathbf{x} \in \mathbb{R}^p} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \in \mathcal{X}, \end{array} \right. \qquad \text{and} \quad d^\star := \max_{\mathbf{x} \in \mathbb{R}^n} d(\lambda).$$

Necessary and sufficient condition

Theorem (Necessary and sufficient optimality condition)

Under Slater's condition (6)*:* $\operatorname{relint}(\mathcal{X}) \cap \{\mathbf{x} \ : \ \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset$*, the* ***KKT condition*** (2)

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

is necessary and sufficient for a point $(\mathbf{x}^\star, \lambda^\star) \in \mathcal{X} \times \mathbb{R}^n$ *being an optimal solution for the primal problem* (1) *and dual problem* (4)*:*

$$f^\star := \left\{ \begin{array}{ll} \min\limits_{\mathbf{x} \in \mathbb{R}^p} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}, \ \mathbf{x} \in \mathcal{X}, \end{array} \right. \qquad \text{and} \quad d^\star := \max_{\mathbf{x} \in \mathbb{R}^n} d(\lambda).$$

Strong duality

- By definition of $f^\star$ and $d^\star$, we always have $\boxed{d^\star \leq f^\star}$ (**weak duality**).
- Under Slater's condition and $\mathcal{X}^\star \neq \emptyset$, we have $\boxed{d^\star = f^\star}$ (**strong duality**).
- Any solution $(\mathbf{x}^\star, \lambda^\star)$ of the KKT condition (2) is also a saddle point.

What happens if Slater's condition does not hold?

Claim

Without Slater's condition, **KKT condition** is only sufficient but not necessary, i.e., if $(\mathbf{x}^\star, \lambda^\star)$ satisfies the KKT condition, then $\mathbf{x}^\star$ is a global solution of (1) but not vice versa.

Example (Violating Slater's condition)

Consider the following **constrained convex** problem:

$$\min_{\mathbf{x}\in\mathbb{R}^2} \{x_1 \ : \ x_2 = 0, x_1^2 - x_2 \leq 0\}$$

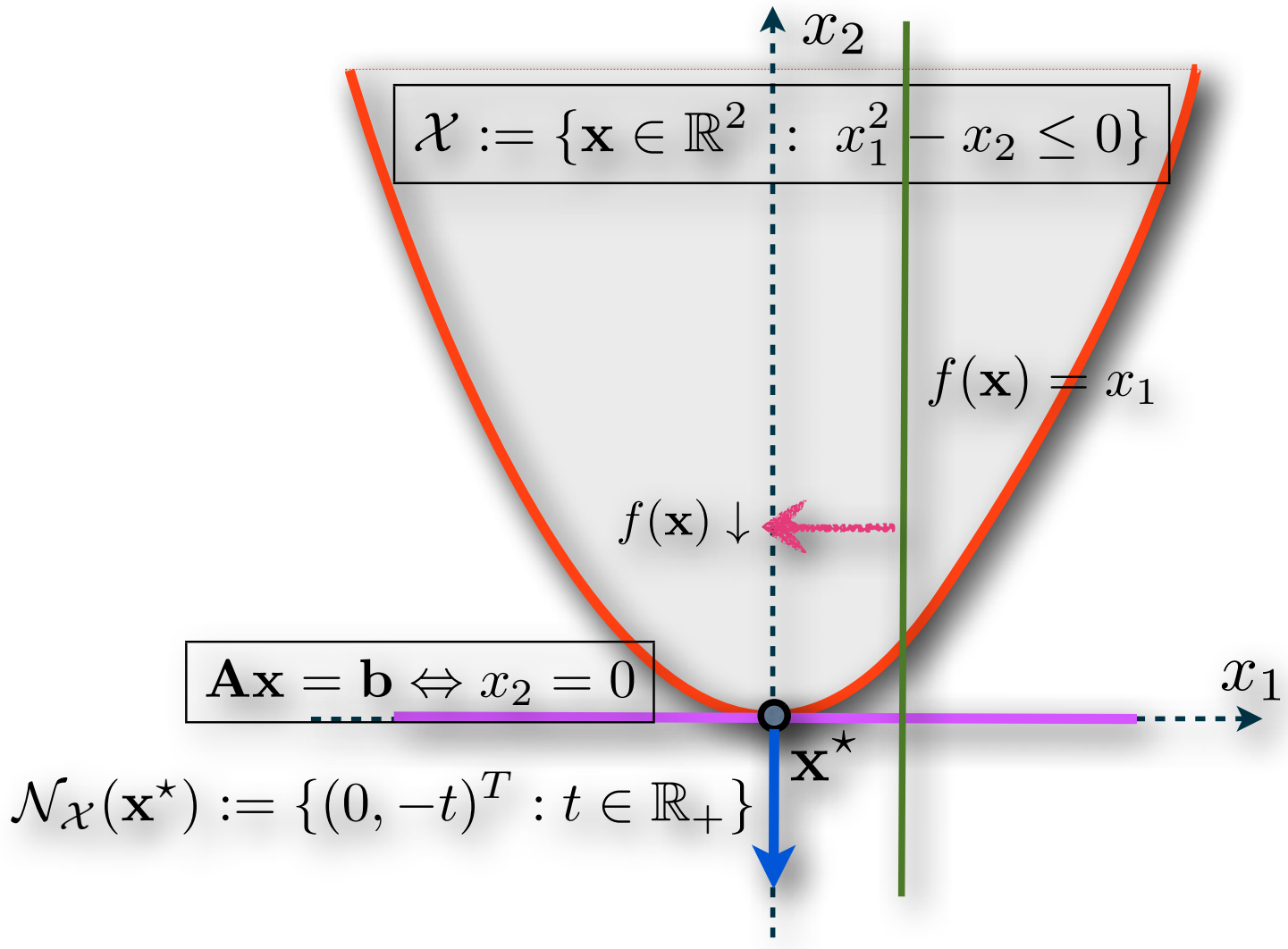
In the setting (1), we have $\mathbf{A} := [0, 1]$, $\mathbf{b} = 0$, $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_1^2 - x_2 \leq 0\}$. The feasible set $\mathcal{D} := \{\mathbf{x} \in \mathbb{R}^2 \ : \ x_2 = 0, x_1^2 - x_2 \leq 0\} = \{(0, 0)^T\}$ contains only one point, which is also the optimal solution of the problem, i.e., $\mathbf{x}^\star := (0, 0)^T$.

In this case, Slater's condition is definitely violated. Let us check the **KKT condition**. Since $\mathcal{N}_{\mathcal{X}}(\mathbf{x}^\star) = \{(0, -t)^T \ : \ t \geq 0\}$, we can write the KKT condition as

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \lambda + \begin{bmatrix} 0 \\ -t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \lambda \in \mathbb{R}, \ t \in \mathbb{R}_+.$$

Since this linear system has no solution due to the first equation $1 = 0$, the **KKT condition** is inconsistent.

Violating Slater's condition



Outline

The proximal way

Establishing correctness

Efficiency considerations

Back to the primal

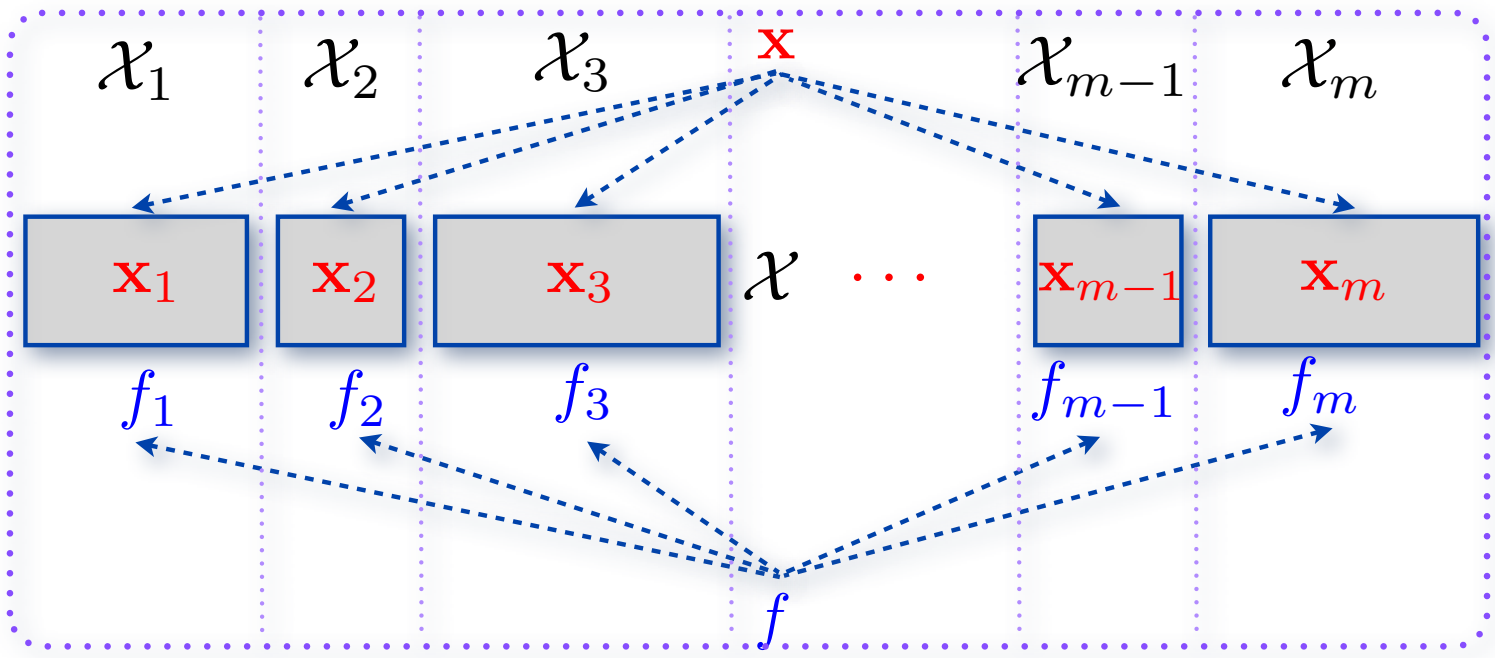
Decomposability

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 $\sum_{i=1}^{m} p_i = p$.



Dual decomposition

An important role of strong duality

- **Strong duality** is a key property in convex optimization, which creates a connection between primal problem (1) and dual problem (4).
- Under Slater's condition, **strong duality** holds, i.e., $f^\star = d^\star$.
- In principle, by solving dual problem (4), we can recover a solution of primal problem (1) and vice versa.

Dual decomposition

An important role of strong duality

- **Strong duality** is a key property in convex optimization, which creates a connection between primal problem (1) and dual problem (4).
- Under Slater's condition, **strong duality** holds, i.e., $f^\star = d^\star$.
- In principle, by solving dual problem (4), we can recover a solution of primal problem (1) and vice versa.

Decomposability is a key property for parallel algorithms

- Under the **decomposable assumption**, the dual function d can be decomposed as

$$d(\lambda) = \sum_{i=1}^{m} d_i(\lambda) - \mathbf{b}^T \lambda.$$

where

$$d_i(\lambda) = \min_{\mathbf{x}_i \in \mathcal{X}_i} \left\{ f_i(\mathbf{x}_i) + \lambda^T \mathbf{A}_i \mathbf{x}_i \right\}, \quad i = 1, \ldots, g.$$

- Evaluating function $d_i(\cdot)$ and its [sub]gradients can be computed in **parallel**

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 $\lambda, \eta \in \text{dom}(d)$ 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 $\lambda, \eta \in \text{dom}(d)$ 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.

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 dual function have Lipschitz gradient?

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

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

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

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

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 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.

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

A simple idea: Apply Nesterov's smoothing [22] to the dual

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

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

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

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

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

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

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

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

Definition (Prox-operator)

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

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.

Outline

The proximal way

Establishing correctness

Efficiency considerations

Back to the primal

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

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)

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

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)

Measuring progress via the gap function

We can define a **gap function** to measure our progress for $\mathbf{z} := (\mathbf{x}, \lambda) \in \mathcal{X} \times \mathbb{R}^n$

$$G(\mathbf{z}) = \underbrace{\max_{\hat{\lambda} \in \mathbb{R}^n} f(\mathbf{x}) + \langle \hat{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle}_{= f(\mathbf{x}) \text{ if } \mathbf{A}\mathbf{x} = \mathbf{b}, \infty \text{ o/w}} - \underbrace{\min_{\hat{\mathbf{x}} \in \mathcal{X}} f(\hat{\mathbf{x}}) + \langle \lambda, \mathbf{A}\hat{\mathbf{x}} - \mathbf{b} \rangle}_{= d(\lambda)} \geq 0$$

- $G(\mathbf{z}^\star) = 0$ 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

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

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

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

Our new technique: Model-based gap reduction MGR (cf., [25])

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

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

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

MGR ties $\bar{\epsilon}$ to ϵ via $f_\beta(\mathbf{x})$

An instance of our primal-dual scheme

The standard scheme ([21])

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

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k \tilde{\lambda}_k \\ \lambda^{k+1} & := \hat{\lambda}^k + \frac{1}{L_{d_\gamma}} \nabla d_\gamma(\hat{\lambda}^k) \\ \tilde{\lambda}_{k+1} & := \lambda_k^* - \frac{1}{\tau_k}(\hat{\lambda}^k - \lambda^{k+1}). \end{cases} \tag{7}$$

Here, L_{d_γ} is the Lipschitz constant of ∇d_γ and $\tau_k \in (0,1)$ is a given momentum term.

Our primal-dual scheme (`http://lions.epfl.ch/decopt`)

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

$$\begin{cases} \hat{\lambda}^k & := (1-\tau_k)\lambda^k + \tau_k \tilde{\lambda}^k \\ \lambda^{k+1} & := \hat{\lambda}^k + \frac{\gamma_{k+1}}{\|\mathbf{A}\|^2}(\mathbf{A}\mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k) - \mathbf{b}) \\ \mathbf{x}^{k+1} & := (1-\tau_k)\mathbf{x}^k + \tau_k \mathbf{x}^*_{\gamma_{k+1}}(\hat{\lambda}^k) \\ \tilde{\lambda}^{k+1} & := \frac{1}{\beta_{k+1}}(\mathbf{A}\mathbf{x}^{k+1} - \mathbf{b}). \end{cases} \tag{8}$$

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

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

An uncertainty relation via MGR ([26, 25])

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 $\tau_k^2 = \Omega\left(\frac{1}{k^2}\right)$ for the smoothed gap.

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

Convergence guarantee

Theorem [26, 25]

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

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

where C_p and C_d are two given positive constants depending on different schemes.

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

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

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

Example: An application of the convergence guarantees

Problem (Consensus optimization)

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right\}$$

Constrained reformulation via a product space trick with $\mathbf{z} := [\mathbf{x}_1, \dots, \mathbf{x}_n]$*:*

$$F^\star := \min_{\mathbf{z}:=[\mathbf{x}_1,\dots,\mathbf{x}_n]\in\mathbb{R}^{np}} \left\{ F(\mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_i) : \mathbf{x}_i - \mathbf{x}_j = 0, (i,j) \in E \right\}$$

for some user-defined graph $\mathcal{G} = (V, E)$ *with vertices* V *and edges* E*.*

Example: An application of the convergence guarantees

Problem (Consensus optimization)

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right\}$$

Constrained reformulation via a product space trick with $\mathbf{z} := [\mathbf{x}_1, \dots, \mathbf{x}_n]$*:*

$$F^\star := \min_{\mathbf{z}:=[\mathbf{x}_1,\dots,\mathbf{x}_n]\in\mathbb{R}^{np}} \left\{ F(\mathbf{z}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}_i) : \mathbf{x}_i - \mathbf{x}_j = 0, (i,j) \in E \right\}$$

for some user-defined graph $\mathcal{G} = (V, E)$ *with vertices* V *and edges* E*.*

Interpretation of the convergence guarantees

By using our algorithm in a decentralized but synchronized fashion, we obtain

$$|F(\mathbf{z}^k) - f^\star| \leq \mathcal{O}(1/k) \quad \text{and} \quad \sum_{(i,j)\in E} \|\mathbf{x}_i^k - \mathbf{x}_j^k\|^2 \leq \mathcal{O}(1/k^2),\ i = 1, \dots, n-1.$$

If f_i's are strongly convex, these rates further improve.

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

The augmented Lagrangian (AL) smoothing

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

1. d_γ now has Lipschitz continuous gradient with $L_{d_\gamma} = \gamma^{-1}$.
2. $\nabla d_\gamma(\lambda) = \mathbf{A}\mathbf{x}_\gamma^*(\lambda) - \mathbf{b}$.
3. $\mathbf{x}_\gamma^*(\lambda)$ can be computed approximately by first-order methods.

Augmented Lagrangian idea: The trade-offs

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

No free lunch: Large γ increases the difficulty of per-iteration time!

Augmented Lagrangian idea: The trade-offs

Theorem (convergence guarantee) [26, 25]

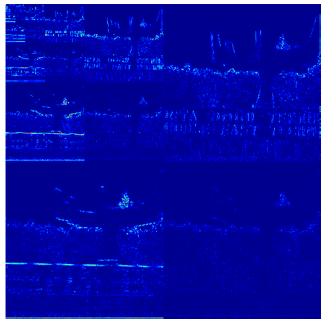
The sequence $\{\mathbf{z}^k\}$ generated by our accelerated scheme satisfies:

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

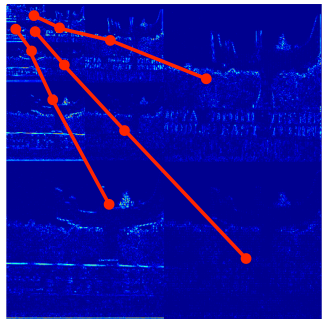
The worst-case iteration complexity: $\mathcal{O}\left(\sqrt{\frac{D_{\Lambda^\star}}{\gamma\epsilon}}\right)$.

- We can increase γ to obtain faster convergence
- However, it becomes more difficult to compute $\mathbf{x}^*_\gamma(\hat{\lambda}^k)$!
- Warm starts help but we need to solve subproblems with increasing accuracy!

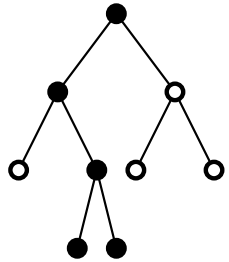
Tree sparsity [19, 10, 2, 31]



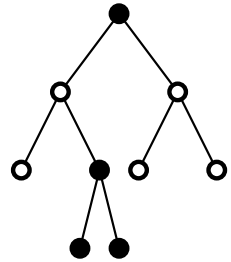
Wavelet coefficients



Wavelet tree

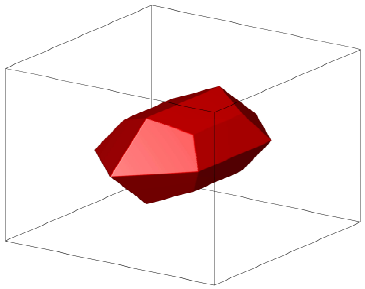


Valid selection of nodes

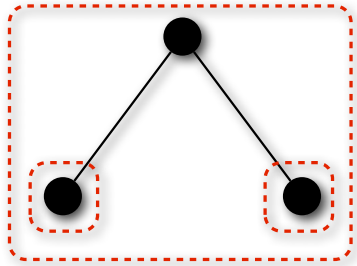


Invalid selection of nodes

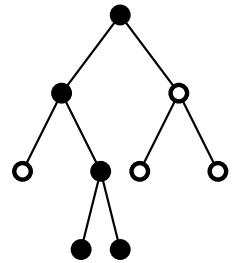
Tree sparsity [19, 10, 2, 31]



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



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



valid selection of nodes

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

Compressive sensing formulation (TU-relax [12])

$$\begin{array}{ll} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) := \sum_{\mathcal{G}_i\in\mathfrak{G}} \|\mathbf{x}_{\mathcal{G}_i}\|_\infty \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{9}$$

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

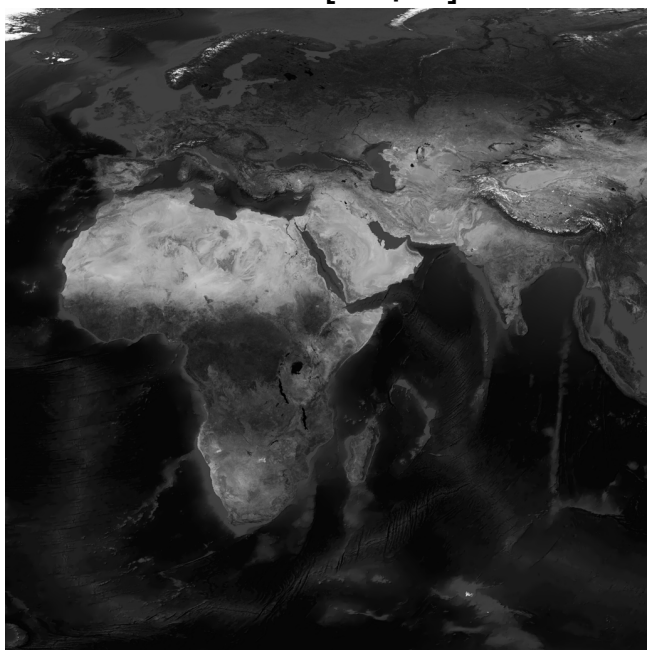
When $g = p$ and $\mathcal{G}_i = \{i\}$, (9) reduces to the well-known basis pursuit (BP):

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{10}$$

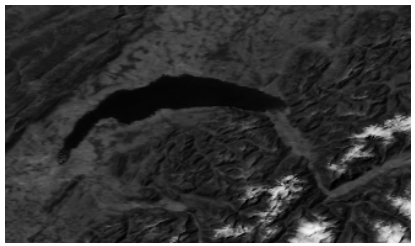
Tree sparsity example: 1:100-compressive sensing

$$(n, p) = (10^7, 10^9)$$

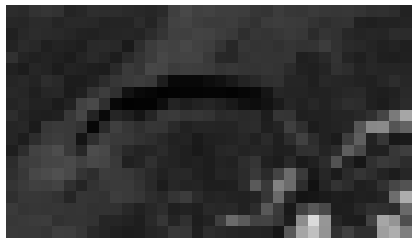
World [1Gpix]



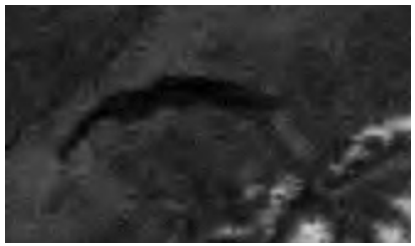
Lac Léman



World [10Mpix]

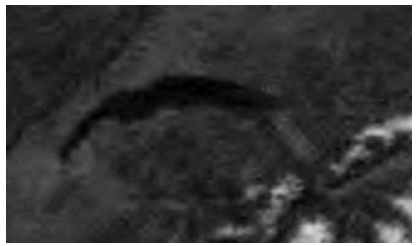


sparse



PNSR $= 31.83$db

tree-sparse



PNSR $= 32.48$db

Sampling: Breaking the coherence barrier [1]

Recovery: Augmented Lagrangian method [26]

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

Tree sparsity example: TV & TU-relax 1:15-compression [25, 1]

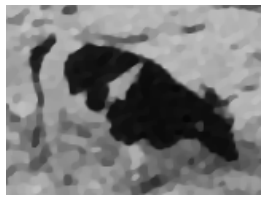
Original tiff image [2048×2048]

Original

BP

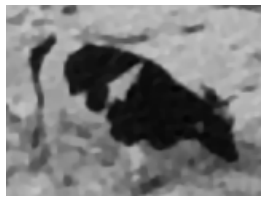
TU-relax

TV

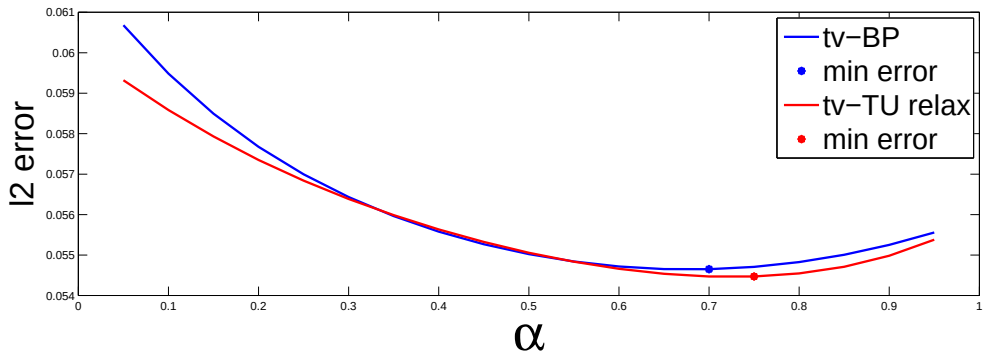


TV with BP

TV with TU-relax



Regularization:



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

[4] 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.

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

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

References II

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

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

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

[10] 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.

[11] 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.

References III

[12] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
In *18th Int. Conf. Artificial Intelligence and Statistics*, 2015.

[13] 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.

[14] 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.

[15] 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.

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

References IV

[17] 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.

[18] 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.

[19] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
J. Mach. Learn. Res., 12:2297–2334, 2011.

[20] 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.

[21] Y. Nesterov.
A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$.
Doklady AN SSSR, 269(translated as Soviet Math. Dokl.):543–547, 1983.

References V

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

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

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

[25] 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.

[26] Q. Tran-Dinh and V. Cevher.
A primal-dual algorithmic framework for constrained convex minimization.
Tech. Report., LIONS, pages 1–54, 2014.

References VI

[27] 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.

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

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

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

[31] 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.