Optimizing a time-data tradeoff via model-based excessive gap

Volkan Cevher
volkan.cevher@epfl.ch

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

Joint work with
Marwa El Halabi, Yen-Huan Li, and Quoc Tran Dinh @ LIONS
Stephen Becker @ UC Boulder
John Burer and Joel Tropp @ Caltech

lions@epfl

MARIE CURIE ACTIONS

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

erc

EPFL

Outline

A time-data conundrum

Constrained convex minimization: The time perspective

Putting it together

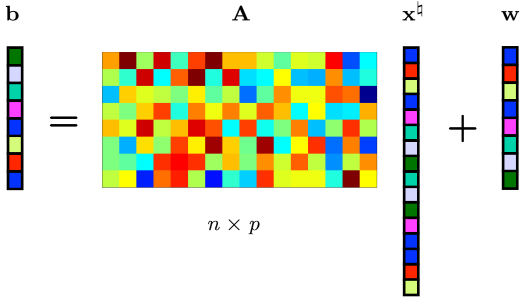
Conclusions

A simple *regression* model

$$\boxed{\mathbf{b}_i = \mathbf{x}^\natural(\mathbf{a}_i) + \mathbf{w}_i}$$

$\mathbf{x}^\natural$: unknown function / hypothesis
$\mathbf{a}_i$: input
$\mathbf{b}_i$: response / output
$\mathbf{w}_i$: perturbations / noise

Linear model:



$$\boxed{\mathbf{b}_i = \mathbf{x}^\natural(\mathbf{a}_i) + \mathbf{w}_i = \langle \mathbf{a}_i, \mathbf{x}^\natural \rangle + \mathbf{w}_i}$$

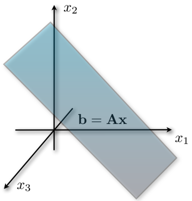
Applications: **Compressive sensing, machine learning, theoretical computer science...**

A simple *regression* model and many *practical* questions

$$\boxed{\mathbf{b}_i = \langle \mathbf{a}_i, \mathbf{x}^\natural \rangle + \mathbf{w}_i}$$

$\mathbf{x}^\natural$: unknown function / hypothesis
$\mathbf{a}_i$: input
$\mathbf{b}_i$: response / output
$\mathbf{w}_i$: perturbations / noise

- Estimation: find $\mathbf{x}^\star$ to minimize $\|\mathbf{x}^\star - \mathbf{x}^\natural\|$
- Prediction: find $\mathbf{x}^\star$ to minimize $\mathcal{L}\left(\mathbf{x}^\star(\mathbf{a}), \mathbf{x}^\natural(\mathbf{a}) + \mathbf{w}\right)$
- Decision: choose $\mathbf{a}_i$ for estimation or prediction



A difficult estimation challenge when $n < p$:

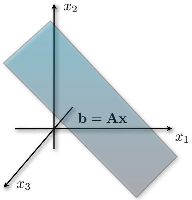
Nullspace (null) of $\mathbf{A}$: $\quad \mathbf{x}^\natural + \boldsymbol{\delta} \to \mathbf{b}, \quad \forall \boldsymbol{\delta} \in \text{null}(\mathbf{A})$

A simple *regression* model and many *practical* questions

$$\boxed{\mathbf{b}_i = \langle \mathbf{a}_i, \mathbf{x}^\natural \rangle + \mathbf{w}_i}$$

$\mathbf{x}^\natural$: unknown function / hypothesis
$\mathbf{a}_i$: input
$\mathbf{b}_i$: response / output
$\mathbf{w}_i$: perturbations / noise

- Estimation: find $\mathbf{x}^\star$ to minimize $\|\mathbf{x}^\star - \mathbf{x}^\natural\|$



A difficult estimation challenge when $n < p$:

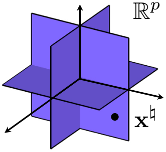
Nullspace (null) of $\mathbf{A}$: $\mathbf{x}^\natural + \boldsymbol{\delta} \to \mathbf{b}, \quad \forall \boldsymbol{\delta} \in \text{null}(\mathbf{A})$

- **Needle in a haystack:** ***We need additional information on $\mathbf{x}^\natural$!***

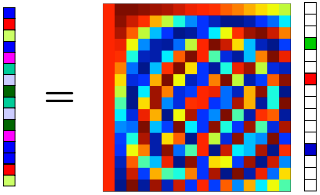
Swiss army knife of signal models

Definition (s-sparse vector)

A vector $\mathbf{x} \in \mathbb{R}^p$ is s-sparse, i.e., $\mathbf{x} \in \Sigma_s$, if it has at most s non-zero entries.



$$\left\|\mathbf{x}^\natural\right\|_0 := |\{i : x_i^\natural \neq 0\}| = s$$

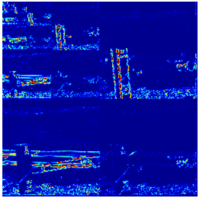


$$\mathbf{y}^\natural = \mathbf{\Psi} \quad \mathbf{x}^\natural$$

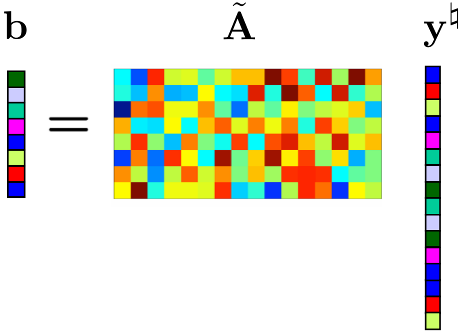
Sparse representations:

$\mathbf{y}^\natural$ has *sparse* transform coefficients $\mathbf{x}^\natural$

- Basis representations $\Psi \in \mathbb{R}^{p \times p}$
  - *Wavelets*, DCT, ...
- Frame representations $\Psi \in \mathbb{R}^{m \times p}$, $m > p$
  - Gabor, curvelets, shearlets, ...
- Other *dictionary* representations...

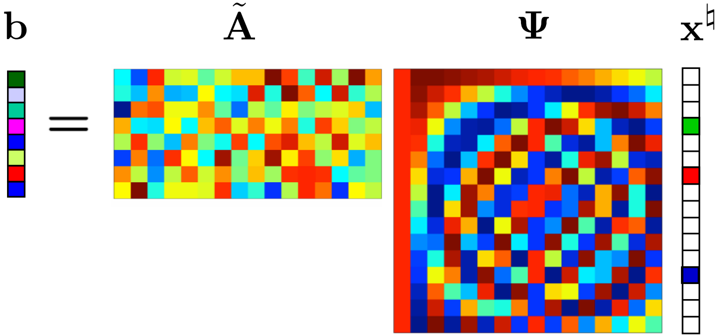


Sparse representations strike back!



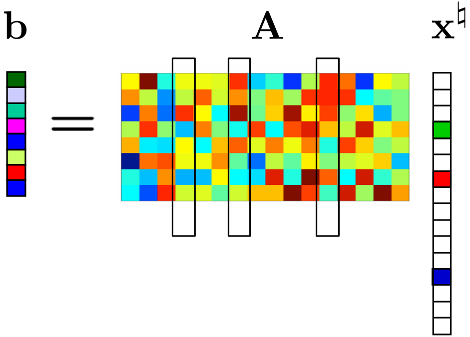
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



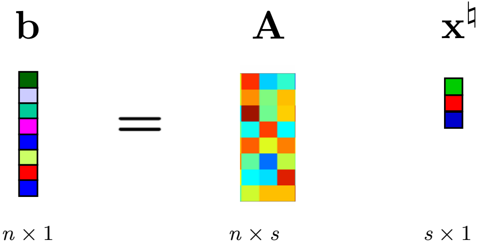
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\boldsymbol{\Psi} \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n < p$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n < p$

Impact: Support restricted columns of $\mathbf{A}$ leads to an *overcomplete* system.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

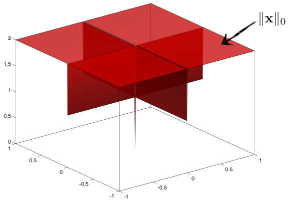
$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

$\|\mathbf{x}\|_0$ over the unit ℓ_∞-ball



Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

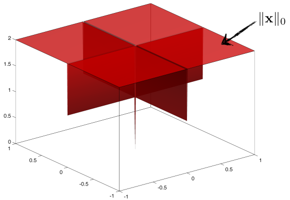
- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Tightest convex relaxation:

$\|\mathbf{x}\|_0^{**}$ is the biconjugate (Fenchel conjugate of Fenchel conjugate)

Fenchel conjugate:
$f^*(\mathbf{y}) := \sup_{\mathbf{x}:\mathrm{dom}(f)} \mathbf{x}^T\mathbf{y} - f(\mathbf{x})$.

$\|\mathbf{x}\|_0$ over the unit ℓ_∞-ball



A technicality: Restrict $\mathbf{x}^\natural \in [-1, 1]^p$.

Enter sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We may consider the estimator with the least number of non-zero entries. That is,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\} \qquad (\mathcal{P}_0)$$

with some $\kappa \geq 0$. If $\kappa = \|\mathbf{w}\|_2$, then $\mathbf{x}^\natural$ is a feasible solution.

$\mathcal{P}_0$ has the following characteristics:

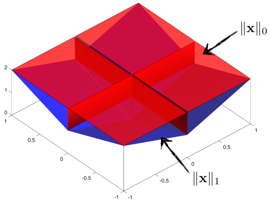
- sample complexity: $\mathcal{O}(s)$
- computational effort: NP-Hard
- stability: No

Tightest convex relaxation:

$\|\mathbf{x}\|_0^{**}$ is the biconjugate (Fenchel conjugate of Fenchel conjugate)

Fenchel conjugate:
$f^*(\mathbf{y}) := \sup_{\mathbf{x}:\text{dom}(f)} \mathbf{x}^T\mathbf{y} - f(\mathbf{x})$.

$\|\mathbf{x}\|_1$ is the convex envelope of $\|\mathbf{x}\|_0$



A technicality: Restrict $\mathbf{x}^\natural \in [-1, 1]^p$.

The role of convexity

A convex candidate solution for $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \le \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \le 1 \right\}. \qquad \text{(SOCP)}$$

Theorem (A **model** recovery guarantee [8])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. Gaussian random variables with zero mean and variances $1/n$. For any $t > 0$ with probability at least $1 - 6\exp\left(-t^2/26\right)$, we have

$$\left\|\mathbf{x}^\star - \mathbf{x}^\natural\right\|_2 \le \left[\frac{2\sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s}}{\sqrt{n} - \sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s} - t}\right]\|\mathbf{w}\|_2 := \varepsilon, \quad \textit{when } \|\mathbf{x}^\natural\|_0 \le s.$$

Observations:

- perfect recovery (i.e., $\varepsilon = 0$) with $n \ge 2s\log(\frac{p}{s}) + \frac{5}{4}s$ whp when $\mathbf{w} = 0$.
- ϵ-accurate solution in $k = \mathcal{O}\left(\sqrt{2p+1}\log(\frac{1}{\epsilon})\right)$ iterations via IPM[1]
  with each iteration requiring the solution of a structured $n \times 2p$ linear system.[2]
- robust to noise.

[1] There is a subtle yet important caveat here that I am sweeping under the carpet!

[2] When $\mathbf{w} = 0$, the IPM complexity (# of iterations × cost per iteration) amounts to $\mathcal{O}(n^2 p^{1.5}\log(\frac{1}{\epsilon}))$.

A Time-Data conundrum — I

A computational dogma

Running time of a learning algorithm increases with the size of the data.

A Time-Data conundrum — I

A computational dogma

Running time of a learning algorithm increases with the size of the data.

- Misaligned goals in the statistical and optimization disciplines

Discipline	Goal	Metric
Optimization	reaching numerical ϵ-accuracy	$\|\mathbf{x}^k - \mathbf{x}^\star\| \leq \epsilon$
Statistics	learning ε-accurate model	$\|\mathbf{x}^\star - \mathbf{x}^\natural\| \leq \varepsilon$

- Main issue: **ϵ and ε are NOT the same but should be treated jointly!**

A Time-Data conundrum — II

A stylized formalization of the time-data tradeoff

The goals of optimization and statistical modeling are tightly connected:

$$\underbrace{\|\mathbf{x}^k - \mathbf{x}^\natural\|}_{\text{learning quality}} \leq \underbrace{\|\mathbf{x}^k - \mathbf{x}^\star\|}_{\epsilon\text{: needs "time" } t(k)} + \underbrace{\|\mathbf{x}^\star - \mathbf{x}^\natural\|}_{\varepsilon\text{: needs "data"} n}$$

$\mathbf{x}^\natural$:	true model in $\mathbb{R}^p$
$\mathbf{x}^\star$:	statistical model estimate
$\mathbf{x}^k$:	numerical solution at iteration k

As the number of data samples n increases

- with a fixed optimization formulation,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \leq 1 \right\}$$

  - numerical methods take longer time t to reach ϵ-accuracy
    - e.g., per-iteration time to solve an $n \times 2p$ linear system
  - statistical model estimates ε become more precise when $\|\mathbf{w}\|_2 = \mathcal{O}(\sqrt{n})$
    - $\varepsilon = \frac{2\sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s}}{\sqrt{n} - \sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s} - t} \|\mathbf{w}\|_2$, with probability $1 - 6\exp(-t^2/26)$.

A Time-Data conundrum — II

A stylized formalization of the time-data tradeoff

The goals of optimization and statistical modeling are tightly connected:

$$\underbrace{\|\mathbf{x}^k - \mathbf{x}^\natural\|}_{\text{learning quality}} \leq \underbrace{\|\mathbf{x}^k - \mathbf{x}^\star\|}_{\epsilon\text{: needs "time" } t(k)} + \underbrace{\|\mathbf{x}^\star - \mathbf{x}^\natural\|}_{\varepsilon\text{: needs "data"} n} \leq \bar{\varepsilon}(t(k), n),$$

$\mathbf{x}^\natural$:	true model in $\mathbb{R}^p$
$\mathbf{x}^\star$:	statistical model estimate
$\mathbf{x}^k$:	numerical solution at iteration k
$\bar{\varepsilon}(t(k), n)$:	actual model precision at time $t(k)$ with n samples

As the number of data samples n increases

- with a fixed optimization formulation,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \leq 1 \right\}$$

  - numerical methods take longer time t to reach ϵ-accuracy
    - e.g., per-iteration time to solve an $n \times 2p$ linear system
  - statistical model estimates ε become more precise when $\|\mathbf{w}\|_2 = \mathcal{O}(\sqrt{n})$
    - $\varepsilon = \dfrac{2\sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s}}{\sqrt{n} - \sqrt{2s\log(\frac{p}{s}) + \frac{5}{4}s} - t}\|\mathbf{w}\|_2$, with probability $1 - 6\exp(-t^2/26)$.

"Time" effort has significant diminishing returns on ε in the underdetermined case*
(cf., [5, 3, 10, 4])

* "Data" effort also exhibits a similar behavior in the overdetermined case when a signal prior is used due to noise!

Data as a computational resource

A stylized formalization of the time-data tradeoff

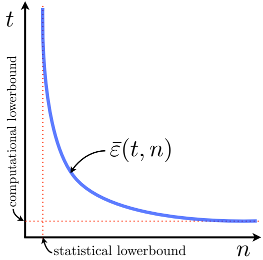
The goals of optimization and statistical modeling are tightly connected:

$$\underbrace{\|\mathbf{x}^{k(t)} - \mathbf{x}^\natural\|}_{\text{learning quality}} \leq \underbrace{\|\mathbf{x}^{k(t)} - \mathbf{x}^\star\|}_{\epsilon\text{: needs ``time'' } t} + \underbrace{\|\mathbf{x}^\star - \mathbf{x}^\natural\|}_{\varepsilon\text{: needs ``data''} n} \leq \bar{\varepsilon}(t,n),$$

$\mathbf{x}^\natural$:	true model in $\mathbb{R}^p$
$\bar{\varepsilon}(t,n)$:	actual model precision at time t with n samples

Rest of the talk:

- primal-dual optimization and sparsity
- a "continuous" time-data tradeoff
  for *underdetermined linear inverse problems*



Outline

Swiss army knife of convex formulations

Our primal problem prototype: A simple mathematical formulation[3]

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

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (1) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

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

Swiss army knife of convex formulations

Our **primal problem** prototype: A simple mathematical formulation[3]

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

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (1) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

Example to keep in mind in the sequel

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

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

Swiss army knife of convex formulations

Our **primal problem** prototype: A simple mathematical formulation[3]

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

- f is a proper, closed and convex function, and $\mathcal{X}$ is a nonempty, closed convex set.
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known.
- An optimal solution $\mathbf{x}^\star$ to (1) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$ and $\mathbf{x}^\star \in \mathcal{X}$.

Example to keep in mind in the sequel

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

Broader context for (1):

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

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

Numerical ϵ-accuracy

Exact vs. approximate solutions

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

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

Definition (ϵ-accurate solutions)

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

$$\begin{cases} |f(\mathbf{x}^\star_\epsilon) - f^\star| \leq \epsilon & \text{(objective residual)}, \\ \|\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}\| \leq \epsilon & \text{(feasibility gap)}, \\ \mathbf{x}^\star_\epsilon \in \mathcal{X} & \text{(exact simple set feasibility).}^4 \end{cases}$$

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

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

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

The optimal solution set

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

Optimality condition

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

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

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

- This is the well-known KKT (Karush-Kuhn-Tucker) condition.
- Any point $(\mathbf{x}^\star, \lambda^\star)$ satisfying (6) is called a KKT point.
- $\mathbf{x}^\star$ is called a stationary point and $\lambda^\star$ is the corresponding multipliers.

Lagrange function and the minimax formulation

We can naturally interpret the optimality condition via a minimax formulation

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

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

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

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (1), we therefore seek the solutions

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

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

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

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

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

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

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (1), we therefore seek the solutions

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

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

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

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

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

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

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

Challenges for the plausible strategy above

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

Finding an optimal solution

A plausible strategy:

To solve the constrained problem (1), we therefore seek the solutions

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

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

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

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

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

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

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

Challenges for the plausible strategy above

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

Efficiency considerations for the dual problem

Subgradient method

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

$$\lambda^{k+1} = \lambda^k + \alpha_k \mathbf{v}^k,$$

where $\mathbf{v}^k \in \partial d(\lambda^k)$ and α_k is the step-size.

Subgradient method for the dual

Assume that the following conditions

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

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

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

Efficiency considerations for the dual problem

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

Subgradient method for the dual

Assume that the following conditions

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

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

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

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

Efficiency considerations for the dual problem

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

Subgradient method for the dual

Assume that the following conditions

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

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

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

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

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

Impact of Lipschitz gradient

(Lipschitz gradient) Let $d(\lambda)$ be a differentiable concave function. $d(\lambda)$ 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

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

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 Lipschitz gradient

(Lipschitz gradient) Let $d(\lambda)$ be a differentiable concave function. $d(\lambda)$ has Lipschitz continuous gradient iff

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

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

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

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

This is NOT the best we can do.

There exists a complexity lower-bound

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

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

Efficiency considerations for the dual problem

Accelerated gradient method

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

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

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

Subgradient method for the dual

Assume that the following conditions

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

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

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

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 Lipschitz gradient

(Lipschitz gradient) Let $d(\lambda)$ be a differentiable concave function. $d(\lambda)$ has Lipschitz continuous gradient iff

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

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

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

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

This is NEARLY the best we can do.

There exists a complexity lower-bound

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

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

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

When can the dual function have Lipschitz gradient?

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

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

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

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

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

When can the dual function have Lipschitz gradient?

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

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

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

Nesterov's smoother [?]

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

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

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

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

Computational efficiency: The key role of the prox-operator

Definition (Prox-operator)

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

Key properties:

- single valued & non-expansive.
- distributes when the primal problem has decomposable structure:

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

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

Computational efficiency: The key role of the prox-operator

Definition (Prox-operator)

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

Key properties:

- single valued & non-expansive.
- distributes when the primal problem has decomposable structure:

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

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

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

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

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

Optimality condition (revisted)

Two equivalent ways of viewing the optimality condition of the primal problem (1)

mixed variational inequality (MVIP) **inclusion**

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

where $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}$ and $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (1).

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

Optimality condition (revisted)

Two equivalent ways of viewing the optimality condition of the primal problem (1)

mixed variational inequality (MVIP) **inclusion**

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

where $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}$ and $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (1).

Measuring progress via the gap function

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

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

Key observations:

- $G(\mathbf{z}) = \underbrace{\max_{\hat{\lambda} \in \mathbb{R}^n} f(\mathbf{x}) + \langle \hat{\lambda}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle}_{= f(\mathbf{x}) \text{ if } \mathbf{A}\mathbf{x} = \mathbf{b}, \infty \text{ o/w}} - \underbrace{\min_{\hat{\mathbf{x}} \in \mathcal{X}} f(\hat{\mathbf{x}}) + \langle \lambda, \mathbf{A}\hat{\mathbf{x}} - \mathbf{b} \rangle}_{= d(\lambda)} \geq 0, \forall \mathbf{z} \in \mathcal{X} \times \mathbb{R}^n$
- $G(\mathbf{z}^\star) = 0$ iff $\mathbf{z}^\star := (\mathbf{x}^\star, \lambda^\star)$ is a primal-dual solution of (1).
- Primal accuracy ϵ and the dual accuracy $\bar{\epsilon}$ can be related via the gap function.

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

A smoothed gap function measuring the excessive primal-dual gap

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

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

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

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

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

A smoothed gap function measuring the excessive primal-dual gap

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

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

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

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

Our algorithm via MEG: model-based excessive gap (cf., [11])

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

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

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

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

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

Key estimates [11, 12]

As a consequence of MEG, we have

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

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

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

Key estimates [11, 12]

As a consequence of MEG, we have

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

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

An uncertainty relation via MEG

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

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

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

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

Convergence guarantee

Recall: Uncertainty relation

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

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

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

Theorem [11, 12]

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

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

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

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

Duality and optimality in constrained convex optimization

Dual problem

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}}\{ \underbrace{\mathcal{L}(\mathbf{x},\lambda)}_{\text{Lagrange function}} := f(\mathbf{x}) + \langle \underbrace{\lambda}_{\text{dual variable}}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle \}. \quad (4)$$

- $\mathbf{x}^\star(\lambda)$ denotes a solution of (4).
- $d(\cdot)$ is concave and generally nonsmooth.

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

$$d^\star := \max_{\lambda\in\mathbb{R}^n} d(\lambda) \quad (5)$$

Duality and optimality in constrained convex optimization

Dual problem

$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}}\{\underbrace{\mathcal{L}(\mathbf{x},\lambda)}_{\text{Lagrange function}} := f(\mathbf{x}) + \langle\underbrace{\lambda}_{\text{dual variable}}, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle\}. \quad (4)$$

- $\mathbf{x}^\star(\lambda)$ denotes a solution of (4).
- $d(\cdot)$ is concave and generally nonsmooth.

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

$$d^\star := \max_{\lambda\in\mathbb{R}^n} d(\lambda) \quad (5)$$

Optimality condition (or KKT condition) of (1)

$$\begin{cases} 0 \in \mathbf{A}^T\lambda^\star + \underbrace{\partial f(\mathbf{x}^\star)}_{\text{subdifferential of } f} + \underbrace{\mathcal{N}_{\mathcal{X}}(\mathbf{x}^\star)}_{\text{normal cone of } \mathcal{X}}, \\ 0 = \mathbf{A}\mathbf{x}^\star - \mathbf{b}, \end{cases} \quad (6)$$

The Slater's condition for (1) becomes:

$$\boxed{\underbrace{\mathrm{relint}(\mathcal{X})}_{\text{relative interior of } \mathcal{X}} \cap \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \quad (7)$$

Optimality condition as a mixed variational inequality (MVIP)

MVIP formulation

- Let $\mathbf{z} := [\mathbf{x}, \lambda]$, $\mathcal{W} := \mathcal{X} \times \mathbb{R}^n$, and $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{b} - \mathbf{A}\mathbf{x} \end{bmatrix}$
- Let $\mathbf{z}^\star := [\mathbf{x}^\star, \lambda^\star]$ be a primal-dual solution of (1).

The **optimality condition** (6) can be written as:

$$\boxed{f(\mathbf{x}) - f(\mathbf{x}^\star) + M(\mathbf{z}^\star)^T(\mathbf{z} - \mathbf{z}^\star) \geq 0, \quad \forall \mathbf{z} \in \mathcal{W}}. \tag{MVIP}$$

Optimality condition as a mixed variational inequality (MVIP)

MVIP formulation

- Let $\mathbf{z} := [\mathbf{x}, \lambda]$, $\mathcal{W} := \mathcal{X} \times \mathbb{R}^n$, and $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{b} - \mathbf{A}\mathbf{x} \end{bmatrix}$
- Let $\mathbf{z}^\star := [\mathbf{x}^\star, \lambda^\star]$ be a primal-dual solution of (1).

The **optimality condition** (6) can be written as:

$$\boxed{f(\mathbf{x}) - f(\mathbf{x}^\star) + M(\mathbf{z}^\star)^T(\mathbf{z} - \mathbf{z}^\star) \geq 0, \quad \forall \mathbf{z} \in \mathcal{W}}. \tag{MVIP}$$

Gap function

We define a **gap function** for (MVIP) as:

$$G(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{W}} \left\{ \mathcal{K}(\mathbf{z}, \hat{\mathbf{z}}) := f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) \right\}. \tag{8}$$

Optimality condition as a mixed variational inequality (MVIP)

MVIP formulation

- Let $\mathbf{z} := [\mathbf{x}, \lambda]$, $\mathcal{W} := \mathcal{X} \times \mathbb{R}^n$, and $M(\mathbf{z}) := \begin{bmatrix} \mathbf{A}^T\lambda \\ \mathbf{b} - \mathbf{A}\mathbf{x} \end{bmatrix}$
- Let $\mathbf{z}^\star := [\mathbf{x}^\star, \lambda^\star]$ be a primal-dual solution of (1).

The **optimality condition** (6) can be written as:

$$\boxed{f(\mathbf{x}) - f(\mathbf{x}^\star) + M(\mathbf{z}^\star)^T(\mathbf{z} - \mathbf{z}^\star) \geq 0, \quad \forall \mathbf{z} \in \mathcal{W}}. \qquad \text{(MVIP)}$$

Gap function

We define a **gap function** for (MVIP) as:

$$G(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{W}} \left\{ \mathcal{K}(\mathbf{z}, \hat{\mathbf{z}}) := f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) \right\}. \qquad (8)$$

Key properties of G

- G is nonnegative for any $\mathbf{z} \in \mathcal{X} \times \mathbb{R}^n$.
- $G(\mathbf{z}^\star) = 0$ iff $\mathbf{z}^\star := [\mathbf{x}^\star, \lambda^\star]$ is a primal-dual solution of (1) and (5).

The algorithmic strategy

Main idea

Finding a **primal-dual solution** $\mathbf{z}^\star$ of (1) and (5) is equivalent to solving

$$G(\mathbf{z}) = 0.$$

Our **strategy** is to design an algorithm such that:

- It generates simultaneously a primal-dual sequence $\{\bar{\mathbf{z}}^k\}_{k\geq 0}$, where $\bar{\mathbf{z}}^k \equiv [\bar{\mathbf{x}}^k, \bar{\lambda}^k]$.
- $G(\bar{\mathbf{z}}^k)$ converges to 0.

The algorithmic strategy

Main idea

Finding a **primal-dual solution** $\mathbf{z}^\star$ of (1) and (5) is equivalent to solving

$$G(\mathbf{z}) = 0.$$

Our **strategy** is to design an algorithm such that:

- It generates simultaneously a primal-dual sequence $\{\bar{\mathbf{z}}^k\}_{k\geq 0}$, where $\bar{\mathbf{z}}^k \equiv [\bar{\mathbf{x}}^k, \bar{\lambda}^k]$.
- $G(\bar{\mathbf{z}}^k)$ converges to 0.

Our approach and contributions

Our approach is mainly inspired by Nesterov's excessive gap technique in [7]:

- Instead of G, we introduce a smoothed gap function $G_{\gamma\beta}$ such that:

$$G_{\gamma_k\beta_k}(\bar{\mathbf{z}}^k) \to G(\bar{\mathbf{z}}^k) \text{ as } \gamma_k\beta_k \to 0^+.$$

- We design different strategies to update the sequence $\{\bar{\mathbf{z}}^k\}$.

Our contributions:

- We estimate optimal convergence rate on $|f(\bar{\mathbf{x}}^k) - f^\star|$ and $\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\|$ separately.
- We cover some well-known algorithms (e.g., accelerated dual gradient, ADMM, PADMM) as concrete instances.

Smoothed gap function

- Let b be a smooth prox-function of $\mathcal{X}$ (strongly convex with parameter $\mu_b = 1$).
- We define ξ the Bregman distance as

$$\xi(\mathbf{x}, \hat{\mathbf{x}}) := b(\mathbf{x}) - b(\hat{\mathbf{x}}) - \nabla b(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}).$$

Smoothed gap function

- Let b be a smooth prox-function of $\mathcal{X}$ (strongly convex with parameter $\mu_b = 1$).
- We define ξ the Bregman distance as

$$\xi(\mathbf{x}, \hat{\mathbf{x}}) := b(\mathbf{x}) - b(\hat{\mathbf{x}}) - \nabla b(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}).$$

Smoothed gap function

Given $\mathbf{S}$ and $\mathbf{x}_c$, we define $\xi_{\gamma\beta}(\mathbf{Sz}, \mathbf{Sz}_c) := \gamma\xi(\mathbf{Sx}, \mathbf{Sx}_c) + (\beta/2)\|\lambda\|_2^2$ and:

$$G_{\gamma\beta}(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{X} \times \mathbb{R}^n} \left\{ f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) - \xi_{\gamma\beta}(\mathbf{S}\hat{\mathbf{z}}, \mathbf{Sz}_c) \right\} \tag{9}$$

the **smoothed gap function** for G.

Smoothed gap function

- Let b be a smooth prox-function of $\mathcal{X}$ (strongly convex with parameter $\mu_b = 1$).
- We define ξ the Bregman distance as

$$\xi(\mathbf{x}, \hat{\mathbf{x}}) := b(\mathbf{x}) - b(\hat{\mathbf{x}}) - \nabla b(\hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}}).$$

Smoothed gap function

Given $\mathbf{S}$ and $\mathbf{x}_c$, we define $\xi_{\gamma\beta}(\mathbf{S}\mathbf{z}, \mathbf{S}\mathbf{z}_c) := \gamma\xi(\mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{x}_c) + (\beta/2)\|\lambda\|_2^2$ and:

$$G_{\gamma\beta}(\mathbf{z}) := \max_{\hat{\mathbf{z}} \in \mathcal{X}\times\mathbb{R}^n} \left\{ f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) - \xi_{\gamma\beta}(\mathbf{S}\hat{\mathbf{z}}, \mathbf{S}\mathbf{z}_c) \right\} \tag{9}$$

the **smoothed gap function** for G.

The choice of $\mathbf{x}_c$ and $\mathbf{S}$

- Bregman distance smoother: $\mathbf{S} := \mathbb{I}$ the identity matrix, and $\mathbf{x}_c$ is fixed at the center point of b (e.g., $\mathbf{x}_c = \arg\min_{\mathbf{x}} b(\mathbf{x})$).
- Augmented Lagrangian smoother: $\mathbf{S} := \mathbf{A}$, and $\mathbf{x}_c \in \mathcal{X}$ such that $\mathbf{A}\mathbf{x}_c = \mathbf{b}$.
  - If $\xi(\mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{x}_c) := \frac{1}{2}\|\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2^k - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{A}_1\mathbf{x}_1^{k+1} + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}\|^2$, we obtain a new ADMM algorithm.
  - If $\xi(\mathbf{S}\mathbf{x}, \mathbf{S}\mathbf{x}_c) := \frac{1}{2}\|\mathbf{x}_1 - \mathbf{g}_1^k\|^2 + \frac{1}{2}\|\mathbf{x}_2 - \mathbf{g}_2^{k+1}\|^2$, where:

$$\begin{cases} g_1^k & := \mathbf{x}_1^k - \|\mathbf{A}_1\|^{-2}\mathbf{A}_1^T(\mathbf{A}_1\mathbf{x}_1^k + \mathbf{A}_2\mathbf{x}_2^k - \mathbf{b}) \\ g_2^{k+1} & := \mathbf{x}_2^k - \|\mathbf{A}_2\|^{-2}\mathbf{A}_2^T(\mathbf{A}_1\mathbf{x}_1^{k+1} + \mathbf{A}_2\mathbf{x}_2^k - \mathbf{b}), \end{cases}$$

    we obtain a new preconditioned ADMM algorithm.

Model-based excessive gap

Model-based excessive gap condition [11, 12]

A **sequence** $\{\bar{\mathbf{z}}^k\}_{k\geq 0} \subset \mathcal{W}$ is said to satisfy the excessive gap condition if:

$$\boxed{G_{k+1}(\bar{\mathbf{z}}^{k+1}) \leq (1-\tau_k)G_k(\bar{\mathbf{z}}^k) - \psi_k} \tag{10}$$

where $G_k(\cdot) := G_{\gamma_k\beta_k}(\cdot)$, $\psi_k \in \mathbb{R}$, $\tau_k \in (0,1)$ and $\gamma_k\beta_{k+1} < \gamma_k\beta_k$ for $k \geq 0$.

Model-based excessive gap

Model-based excessive gap condition [11, 12]

A **sequence** $\{\bar{\mathbf{z}}^k\}_{k\geq 0} \subset \mathcal{W}$ is said to satisfy the excessive gap condition if:

$$\boxed{G_{k+1}(\bar{\mathbf{z}}^{k+1}) \leq (1-\tau_k)G_k(\bar{\mathbf{z}}^k) - \psi_k} \tag{10}$$

where $G_k(\cdot) := G_{\gamma_k\beta_k}(\cdot)$, $\psi_k \in \mathbb{R}$, $\tau_k \in (0,1)$ and $\gamma_k\beta_{k+1} < \gamma_k\beta_k$ for $k \geq 0$.

How to use this model-based excessive gap condition?

Generate a primal-dual sequence $\{\bar{\mathbf{z}}^k\}_{k\geq 0}$ with $\bar{\mathbf{z}}^k := (\bar{\mathbf{x}}^k, \bar{\lambda}^k)$ such that

$$G_{\gamma_k\beta_k}(\bar{\mathbf{z}}^k) \to 0^+$$

by controlling γ_k and $\beta_k \to 0^+$.

Key observation:

- When γ_k and β_k go to zero, we have $\boxed{G_{\gamma_k\beta_k}(\cdot) \to G(\cdot)}$.
- **Consequence**: $\boxed{G(\mathbf{z}^k) \to 0^+ \quad \Rightarrow \quad \bar{\mathbf{z}}^k \to \mathbf{z}^\star = (\mathbf{x}^\star, \lambda^\star)}$ (primal-dual solution).

Key estimate and the evaluation of $G_{\gamma\beta}$

Theorem (Bounds on the objective residual and primal feasibility)

Assume that $\{\bar{\mathbf{z}}^k\}_{k\geq 0}$ *is a sequence satisfying* (10). *Then*

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^\star| & \leq \max\left\{\gamma_k D_{\mathcal{X}}, \left(2\beta_k D_{\Lambda^\star} + \sqrt{2\gamma_k\beta_k D_{\mathcal{X}}}\right) D_{\Lambda^\star}\right\}, \\ \|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| & \leq 2\beta_k D_{\Lambda^\star} + \sqrt{2\gamma_k\beta_k D_{\mathcal{X}}}, \end{cases} \tag{11}$$

where

- $D_{\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} \xi(\mathbf{x}, \mathbf{x}_c)$ *the prox-diameter of* $\mathcal{X}$
- $D_{\Lambda^\star} := \min\{\|\lambda^\star\| \ : \ \lambda^\star \in \Lambda^\star\}$ *the* ***norm*** *of minimum norm solutions of* (5).

Key estimate and the evaluation of $G_{\gamma\beta}$

Theorem (Bounds on the objective residual and primal feasibility)

Assume that $\{\bar{\mathbf{z}}^k\}_{k\geq 0}$ *is a sequence satisfying* (10). *Then*

$$\begin{cases} |f(\bar{\mathbf{x}}^k) - f^\star| & \leq \max\left\{\gamma_k D_{\mathcal{X}}, \left(2\beta_k D_{\Lambda^\star} + \sqrt{2\gamma_k\beta_k D_{\mathcal{X}}}\right) D_{\Lambda^\star}\right\}, \\ \|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| & \leq 2\beta_k D_{\Lambda^\star} + \sqrt{2\gamma_k\beta_k D_{\mathcal{X}}}, \end{cases} \tag{11}$$

where

- $D_{\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} \xi(\mathbf{x}, \mathbf{x}_c)$ *the prox-diameter of* $\mathcal{X}$
- $D_{\Lambda^\star} := \min\{\|\lambda^\star\| \ : \ \lambda^\star \in \Lambda^\star\}$ *the **norm** of minimum norm solutions of* (5).

Evaluation of $G_{\gamma\beta}$

The evaluation of $G_{\gamma\beta}$ requires to solve:

$$G_{\gamma\beta}(\mathbf{z}) := \max_{\hat{\mathbf{z}}\in\mathcal{X}\times\mathbb{R}^n} \left\{f(\mathbf{x}) - f(\hat{\mathbf{x}}) + M(\mathbf{z})^T(\mathbf{z} - \hat{\mathbf{z}}) - \xi_{\gamma\beta}(\mathbf{S}\hat{\mathbf{z}}, \mathbf{S}\mathbf{z}_c)\right\}.$$

The solution $\mathbf{z}^\star_{\gamma\beta}(\mathbf{z}) := (\mathbf{x}^\star_\gamma(\lambda), \lambda^\star_\beta(\mathbf{x}))$ of this problem is given as:

$$\begin{cases} \mathbf{x}^\star_\gamma(\lambda) & := \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{f(\mathbf{x}) + (\mathbf{A}^T\lambda)^T\mathbf{x} + \gamma\xi(\mathbf{S}_x\mathbf{x}, \mathbf{S}_x\mathbf{x}_c)\right\} \\ \lambda^\star_\beta(\mathbf{x}) & := \beta^{-1}(\mathbf{A}\mathbf{x} - \mathbf{b}). \end{cases} \tag{12}$$

The primal-dual scheme: two ingredients

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

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

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

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

The primal-dual scheme: two ingredients

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

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

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

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

Update parameters

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

$$\gamma_{k+1} := (1 - c_k\tau_k)\gamma_k \quad \text{and} \quad \beta_{k+1} = (1-\tau_k)\beta_k \tag{13}$$

The parameter τ_k is updated as:

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

Convergence guarantee

Theorem (Convergence [11, 12])

Let $\{\bar{\mathbf{z}}^k\}$ *be generated by our* ***primal-dual*** *algorithm (1P2D). Then:*

a) *If* $\mathbf{S} = \mathbf{A}$ *(augmented Lagrangian smoother),* $\gamma_0 := 1$ *and* $c_k = 0$*, then*

$$\begin{cases} & \|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| & \leq \frac{D_{\Lambda^\star}}{(k+1)^2}, \\ -(1/2)\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\|^2 - D_{\Lambda^\star}\|\mathbf{A}\bar{\mathbf{x}}^k - \mathbf{b}\| & \leq f(\bar{\mathbf{x}}^k) - f^\star & \leq 0. \end{cases}$$

b) *If* $\mathbf{S} = \mathbb{I}$ *(Bregman smoother),* $\gamma_0 := \frac{2\sqrt{2}\|\mathbf{A}\|}{K+1}$ *and* $c_k = 0$ *for all* $0 \leq k \leq K$*, then*

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

c) *If* $\mathbf{S} = \mathbb{I}$ *(Bregman smoother) and* f *is strongly convex with* $\mu_f > 0$*, then*

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

Sample complexity analysis

Convex optimization formulation for the estimator

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}\},$$

where $f : \mathbb{R}^p \to \mathbb{R} \cup \{-\infty, \infty\}$ is a convex function.

Sample complexity

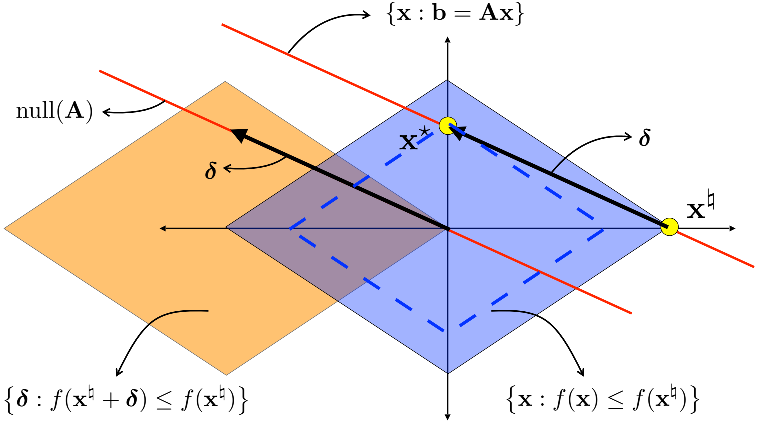
Assume that $A \in \mathbb{R}^{n\times p}$ is a matrix of independent identically distributed (i.i.d.) standard Gaussian random variables.

What is the minimum number of samples n such that $\mathbf{x}^\star = \mathbf{x}^\natural$ with high probability?

Characterization of the error vector

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

Define the error vector $\boldsymbol{\delta} := \mathbf{x}^\star - \mathbf{x}^\natural$.

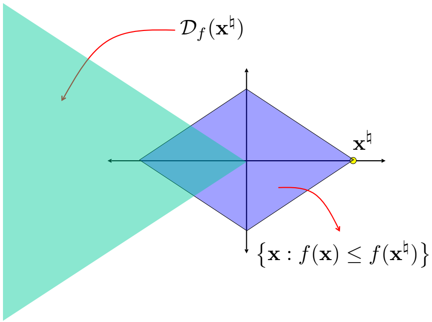


Descent cone

Definition (Descent cone)

Let $f : \mathbb{R}^p \to \mathbb{R} \cup \{-\infty, \infty\}$ be a proper lower-semicontinuous function. The descent cone of f at $\mathbf{x}^\natural$ is defined as

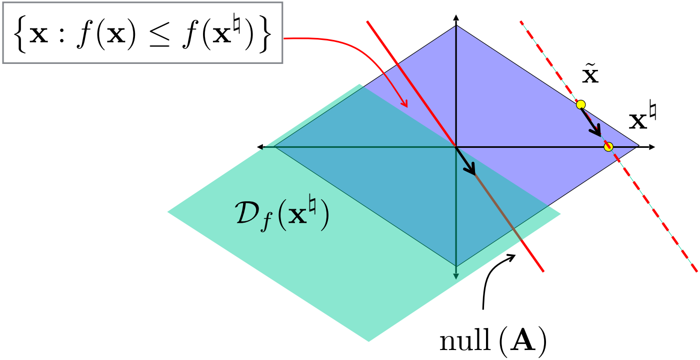
$$\mathcal{D}_f(\mathbf{x}^\natural) := \operatorname{cone}\left(\left\{\mathbf{x} : f(\mathbf{x}^\natural + \mathbf{x}) \leq f(\mathbf{x}^\natural)\right\}\right).$$



Condition for exact recovery in the *noiseless* case

Proposition (Condition for exact recovery)

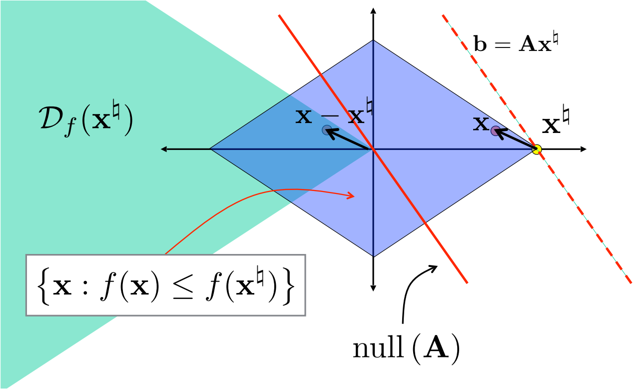
We have successful recovery, i.e., $\boldsymbol{\delta} := \mathbf{x}^\star - \mathbf{x}^\natural = 0$ *with* $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}\}$, *if and only if* $\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}$.



Condition for exact recovery in the *noiseless* case

Proposition (Condition for exact recovery)

We have successful recovery, i.e., $\boldsymbol{\delta} := \mathbf{x}^\star - \mathbf{x}^\natural = 0$ *with* $\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}\}$, *if and only if* $\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}$.



Statistical dimension and approximate kinematic formula

Now we have

$$\mathbb{P}\left\{\mathbf{x}^\star = \mathbf{x}^\natural\right\} = \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{D}_f(\mathbf{x}^\natural) = \{0\}\right\}.$$

Definition (Statistical dimension [1][5])

Let $\mathcal{C} \subseteq \mathbb{R}^p$ be a closed convex cone. The statistical dimension of $\mathcal{C}$ is defined as

$$d(\mathcal{C}) := \mathbb{E}\left[\|\mathrm{proj}_{\mathcal{C}}(\mathbf{g})\|_2^2\right].$$

Theorem (Approximate kinematic formula [1])

Let $A \in \mathbb{R}^{n\times p}$, $n < p$, be a matrix of i.i.d. standard Gaussian random variables, and let $\mathcal{C} \subseteq \mathbb{R}^p$ be a closed convex cone. Let $\eta \in (0,1)$ Then

$$\begin{aligned} n \geq d(\mathcal{C}) + c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{C} = \{0\}\right\} \geq 1-\eta; \\ n \leq d(\mathcal{C}) - c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left\{\mathrm{null}(\mathbf{A}) \cap \mathcal{C} = \{0\}\right\} \leq \eta, \end{aligned}$$

where $c_\eta := \sqrt{8\log(4/\eta)}$.

[5] The statistical dimension is closely related to the Gaussian complexity [2], Gaussian width [6], mean width [13], and Gaussian squared complexity [5].

Probability of exact recovery

Corollary

For any $\eta \in (0,1)$,

$$n \geq d(\mathcal{D}_f(\mathbf{x}^\natural)) + c_\eta \sqrt{p} \quad \Rightarrow \quad \mathbb{P}\left\{\mathbf{x}^\star = \mathbf{x}^\natural\right\} \geq 1 - \eta;$$
$$n \leq d(\mathcal{D}_f(\mathbf{x}^\natural)) - c_\eta \sqrt{p} \quad \Rightarrow \quad \mathbb{P}\left\{\mathbf{x}^\star = \mathbf{x}^\natural\right\} \leq \eta,$$

where $c_\eta := \sqrt{8\log(4/\eta)}$.

- There is a *phase transition* at $n \approx d(\mathcal{D}_f(\mathbf{x}^\natural))$.

Examples ([1, 9])

- Let $f(\mathbf{x}) := \|\mathbf{x}\|_1$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ be s-sparse. Then $d(\mathcal{D}_f(\mathbf{x}^\natural)) \leq 2s\log(p/s) + (5/4)s$.
- Let $f(\mathbf{x}) := \|\mathbf{X}\|_*$, and let $\mathbf{X}^\natural \in \mathbb{R}^{p \times p}$ of rank r. Then $d(\mathcal{D}_f(\mathbf{x}^\natural)) \leq 3r(2p - r)$.
- Let $\mathfrak{G} \subset 2^{\{1,\ldots,p\}}$ be a set of non-overlapping groups. Let $f(\mathbf{x}) := \sum_{\mathcal{G} \in \mathfrak{G}} \|\mathbf{x}_\mathcal{G}\|_2$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ be k-group sparse. Denoting B to be the maximal group size, we have $d(\mathcal{D}_f(\mathbf{x}^\natural)) \leq \left(\sqrt{2\log\left(|\mathfrak{G}| - k\right)} + \sqrt{B}\right)^2 k + Bk$.

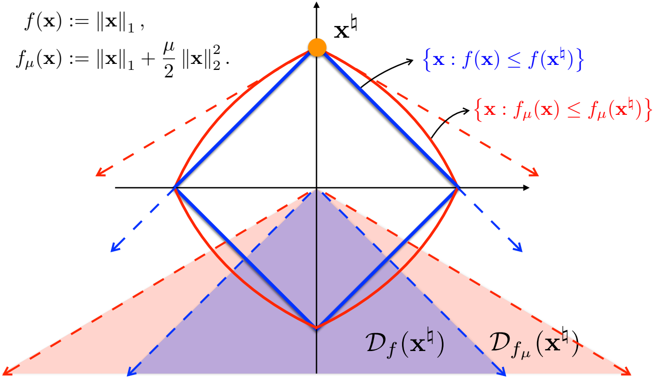
Outline

Smoothing increases the statistical dimension

Key properties of the statistical dimension [1]

- The statistical dimension is invariant under unitary transformations (rotations).
- Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be closed convex cones. If $\mathcal{C}_1 \subseteq \mathcal{C}_2$, then $d(\mathcal{C}_1) \leq d(\mathcal{C}_2)$.

The larger the statistical dimension is, the more number of observations is required.



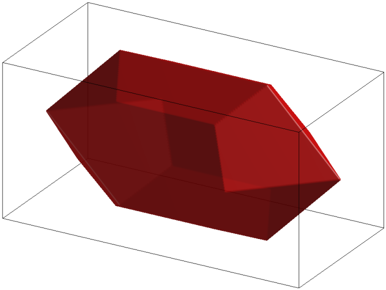
Smoothing increases the statistical dimension

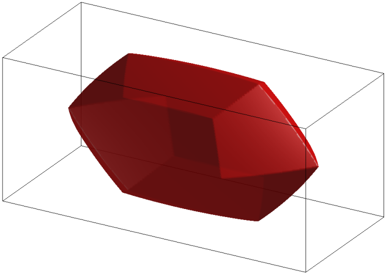
Take the following group norm as an example:

$$f(\mathbf{x}) := \sum_{\mathcal{G} \in \mathfrak{G}_H} \|\mathbf{x}_{\mathcal{G}}\|_\infty .$$

Define

$$f_\mu(\mathbf{x}) := \sum_{\mathcal{G} \in \mathfrak{G}_H} \left(\|\mathbf{x}_{\mathcal{G}}\|_\infty + \frac{\mu}{2} \|\mathbf{x}_{\mathcal{G}}\|_2^2 \right) .$$





unit $f(\mathbf{x})$ ball vs. unit $f_\mu(\mathbf{x})$ ball

Calculation of $d\left(\mathcal{D}_f\left(\mathbf{x}^\natural\right)\right)$ and $d\left(\mathcal{D}_{f_\mu}\left(\mathbf{x}^\natural\right)\right)$

Lemma ([1])

Let f be a proper lower-semicontinuous convex function, and let $\mathbf{x} \in$ dom(f). We have

$$d\left(\mathcal{D}_f\left(\mathbf{x}\right)\right) \leq \inf_{\tau>0} \mathbb{E}\left[\mathrm{dist}^2\left(\mathbf{g}, \tau\partial f(\mathbf{x})\right)\right],$$

where $\mathbf{g}$ is a vector of i.i.d. standard Gaussian random variables.

The upper bounds on $d\left(\mathcal{D}_f\left(\mathbf{x}^\natural\right)\right)$ and $d\left(\mathcal{D}_{f_\mu}\left(\mathbf{x}^\natural\right)\right)$ can be derived based on above.

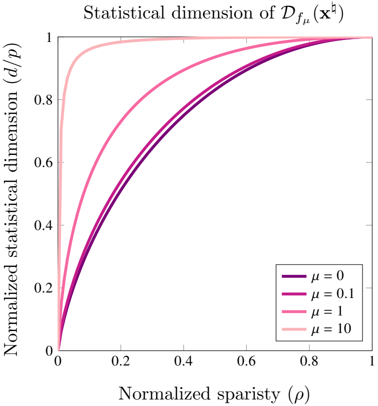
Proposition ([4])

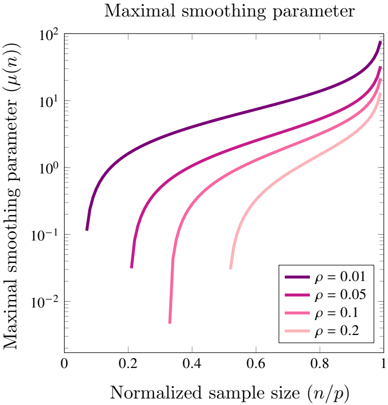
Let $\mathbf{x}^\natural$ be an s-sparse vector. We have

$$d\left(\mathcal{D}_{f_\mu}\left(\mathbf{x}^\natural\right)\right) \leq \inf_{\tau>0} \left\{ s(1+\tau^2) + 2\mu f_\mu(\mathbf{x}^\natural)\tau^2 + (p-s)\sqrt{\frac{2}{\pi}}\int_\tau^\infty (u-\tau)^2 e^{-u^2/2}\, du \right\}.$$

Note that $f = f_\mu|_{\mu=0}$.

Numerical results for the statistical dimension and $\mu(n)$





Smoothing decreases the computational cost

Consider the estimator,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f_\mu(\mathbf{x}) : \mathbf{b} = \mathbf{A}\mathbf{x}, \|\mathbf{x}\|_\infty \leq \|\mathbf{x}^\natural\|_\infty \right\}, \quad \mu \in [0,\infty).$$

Proposition ([4])

Let $\mu > 0$ and $f(\mathbf{x}) = \|\mathbf{x}\|_1$. Consider solving (14) with our primal-dual method. The output after the k-th iteration, $\mathbf{x}^k$, satisfies

$$\left\|\mathbf{x}^\star - \mathbf{x}^k\right\|_2 \leq \frac{4p\kappa(\mathbf{A})\left[\rho(1+\mu\|\mathbf{x}^\star\|_\infty)^2 + (1-\rho)\right]}{\mu k} \propto \left.\frac{1}{\mu k}\right|_{\rho \ll 1},$$

where $\rho := s/p$, s being the number of non-zero entries in $\mathbf{x}^\star$, and $\kappa(\mathbf{A})$ denotes the restricted condition number of $\mathbf{A}$.

Observation:

- When $\rho \ll 1$, the number of iterations k required to achieve the error bound $\left\|\mathbf{x}^\star - \mathbf{x}^k\right\|_2 \leq \varepsilon$ for a fixed $\epsilon > 0$, is proportional to $1/(\mu\varepsilon)$.

Time-data tradeoff

Define the maximal smoothing parameter

$$\mu(n) := \arg\max_{\mu>0} \left\{ \mu : d\left(\mathcal{D}_{f_\mu}(\mathbf{x}^\natural)\right) \leq n \right\}.$$

Consider the "conservative" estimator in probability,

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f_\mu(\mathbf{x})\big|_{\mu=\frac{1}{4}\mu(n)} : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}.$$

Corollary

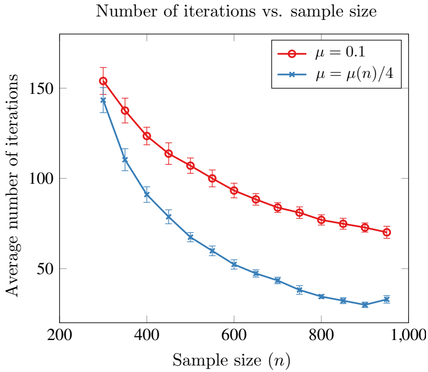
Let $\rho := s/p \ll 1$. Then we have, with high probability, $\mathbf{x}^\star = \mathbf{x}^\natural$, and

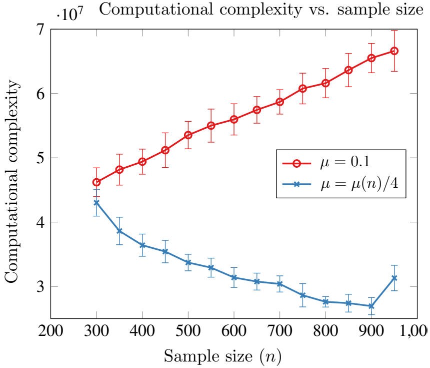
$$\left\|\mathbf{x}^\natural - \mathbf{x}^k\right\|_2 \propto \frac{1}{\mu(n)k}.$$

Therefore, to achieve the error bound, $\left\|\mathbf{x}^\natural - \mathbf{x}^k\right\|_2 \leq \varepsilon$ for a fixed $\varepsilon > 0$, it suffices to choose

$$k = O\left(\frac{1}{\mu(n)}\right).$$

A numerical result for the time-data tradeoff





Outline

Conclusions

- When n is large, we can exploit excess samples beyond the statistical dimension
  - to decrease estimation error / statistical risk (forthcoming)
  - to decrease computational cost
  - to trade off between the two
- Exploring the tradeoff requires a unified analysis in both optimization & statistics
  - convexity acts as a catalyst towards this direction
- Our contributions:
  - a (generative) TU view of sparsity with tightness guarantees
  - model-based excessive gap for construction and analysis of primal-dual algorithms
  - statistical dimension calculations to establish the tradeoff

References I

[1] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp.
Living on the edge: Phase transitions in convex programs with random data.
Inf. Inference, 3:224–294, 2014.
arXiv:1303.6672v2 [cs.IT].

[2] Peter L. Barlett and Shahar Mendelson.
Rademacher and Gaussian complexities: Risk bounds and structural results.
J. Mach. Learn. Res., 3, 2002.

[3] Léon Bottou and Oliver Bousquet.
The tradeoffs of large scale learning.
In *Advances in Neural Information Processing Systems*, 2007.

[4] John J. Bruer, Joel A. Tropp, Volkan Cevher, and Stephen R. Becker.
Time-data tradeoffs by aggressive smoothing.
In *Conference of Neural Information Processing Systems (NIPS)*, 2014.

[5] Venkat Chandrasekaran and Michael I. Jordan.
Computational and statistical tradeoffs via convex relaxation.
Proc. Natl. Acad. Sci., 110(13):E1181–E1190, 2013.

[6] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky.
The convex geometry of linear inverse problems.
Found. Comput. Math., 12:805–849, 2012.

References II

[7] Y. Nesterov.
Excessive gap technique in nonsmooth convex minimization.
SIAM J. Optimization, 16(1):235–249, 2005.

[8] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi.
Simple bounds for noisy linear inverse problems with exact side information.
2013.
arXiv:1312.0641v2 [cs.IT].

[9] Nikhil Rao, Benjamin Recht, and Robert Nowak.
Universal measurement bounds for structured sparse signal recovery.
In *Advances in Neural Information Processing Systems*, 2012.

[10] Shai Shalev-Shwartz and Nathan Srebro.
SVM optimization: inverse dependence on training set size.
In *Proceedings of the 25th international conference on Machine learning*, pages 928–935. ACM, 2008.

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

References III

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

[13] Roman Vershynin.
Estimation in high dimensions: A geometric perspective.
2014.
arXiv:1405.5103v1 [math.ST].