Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

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

EE-556 (Fall 2014)

lions@epfl

License Information for Mathematics of Data Slides

- This lecture
  1. Deficiency of smooth models
  2. Nonsmooth models via atomic norms
  3. Statistical analysis of the basis pursuit denoising estimator
  4. Restricted isometry property and its implications
  5. Lasso, regularized least-squares and their relations to basis pursuit denoising
  6. Selecting a good regularization coefficient
- Next lecture
  1. Unconstrained, non-smooth composite minimization
  2. Convergence and convergence rate characterization of various approaches

Recommended Reading

- V. Chandrasekaran, *et al.*, "The convex geometry of linear inverse problems," *Found. Comput. Math.*, vol. 12, pp. 805–849, 2012.
- J. A. Tropp, "Convex recovery of a structured signal from independent random linear measurements," 2014, arXiv:1405.1102v1 [cs.IT].
- Chapter 2 & Chapter 6 in S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.
- Chapter 7 in T. Hastie *et al.*, *The Elements of Statistical Learning*. Springer, 2009.

Motivation

Motivation

Nonsmooth convex models can help improve the statistical accuracy in estimation.

To this end, this lecture characterizes an important class of nonsmooth optimization models and establishes rigorous estimation guarantees.

Nonsmooth convex models can also lead to *regularized* convex formulations for estimation. This lecture also studies principled approaches to select the regularization parameter.

Deficiency of smooth models

Recall the practical performance of an estimator $\hat{\mathbf{x}}$.

Practical performance

Denote the numerical approximation by $\mathbf{x}_\epsilon^\star$. The practical performance is determined by

$$\left\|\mathbf{x}_\epsilon^\star - \mathbf{x}^\natural\right\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} .$$

Sometimes *non-smooth* convex models of $\mathbf{x}^\natural$ can help *reduce the statistical error*.

Example: Least-squares estimation in the linear model

Recall the linear model and the LS estimator.

LS estimation in the linear model

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes the unknown noise.

The LS estimator for $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b}$ is defined as

$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}.$$

- If $\mathbf{A}$ has full column rank, $\hat{\mathbf{x}}_{\mathsf{LS}} = \mathbf{A}^\dagger \mathbf{b}$ is uniquely defined.
- If $n < p$, then $\mathbf{A}$ cannot have full column rank, and we can only conclude that $\hat{\mathbf{x}}_{\mathsf{LS}} \in \left\{ \mathbf{A}^\dagger \mathbf{b} + \mathbf{h} : \mathbf{h} \in \operatorname{null}(\mathbf{A}) \right\}$.

Observation: The estimation error $\left\| \hat{\mathbf{x}}_{\mathsf{LS}} - \mathbf{x}^\natural \right\|_2$ can be *arbitrarily large*!

A candidate solution

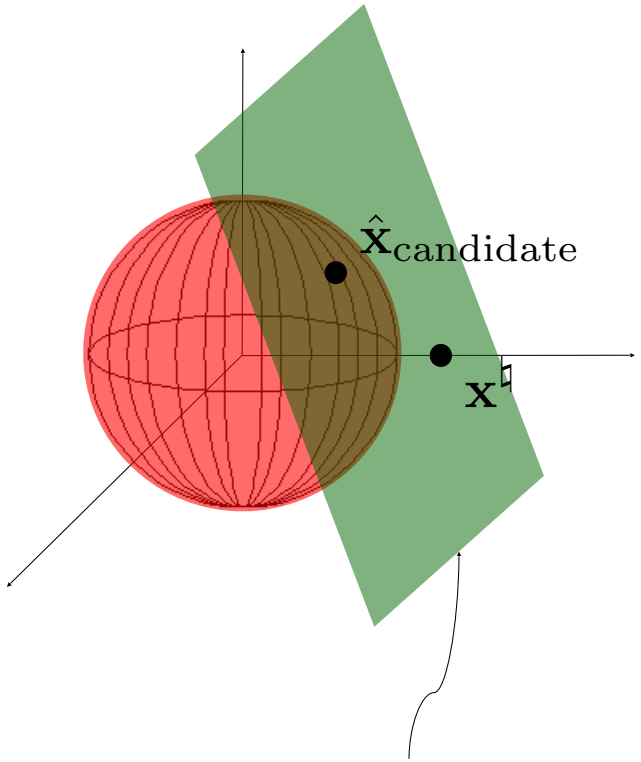
Observation: When $\mathbf{A}$ has full column rank and $\mathbf{w} = 0$,

$$\hat{\mathbf{x}}_{\mathsf{LS}} = \mathbf{A}^{\dagger}\mathbf{b} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_2^2 : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}.$$

Can we use $\hat{\mathbf{x}}_{\mathrm{candidate}} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_2^2 : \mathbf{b} = \mathbf{A}\mathbf{x} \right\} = \mathbf{A}^{\dagger}\mathbf{b}$ even when $n < p$?

Geometry due to the candidate estimator in the noiseless case

$$\hat{\mathbf{x}}_{\text{candidate}} = \mathbf{A}^{\dagger}\mathbf{b} = \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_2^2 : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}.$$



A candidate solution contd.

Proposition ([23])

Suppose that $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a matrix of i.i.d. standard Gaussian random variables, and* $\mathbf{w} = \mathbf{0}$. *We have*

$$(1-\epsilon)\left(1-\frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2 \leq \left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2 \leq (1-\epsilon)^{-1}\left(1-\frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2$$

with probability at least $1 - 2\exp\left[-(1/4)(p-n)\epsilon^2\right] - 2\exp\left[-(1/4)p\epsilon^2\right]$, *for all* $\epsilon > 0$ *and* $\mathbf{x}^\natural \in \mathbb{R}^p$.

Observation: The estimation error may *not* diminish unless n is very close to p.

Intuition: The relation $n < p$ means that the dimension of the sample $\mathbf{b}$ exceeds the number of unknown variables in $\mathbf{x}^\natural$ to be solved.

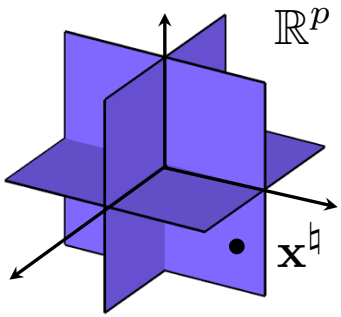
Impact: It is impossible to estimate $\mathbf{x}^\natural$ accurately using $\hat{\mathbf{x}}_{\text{candidate}}$ when $n \ll p$ even if $\mathbf{w} = \mathbf{0}$.

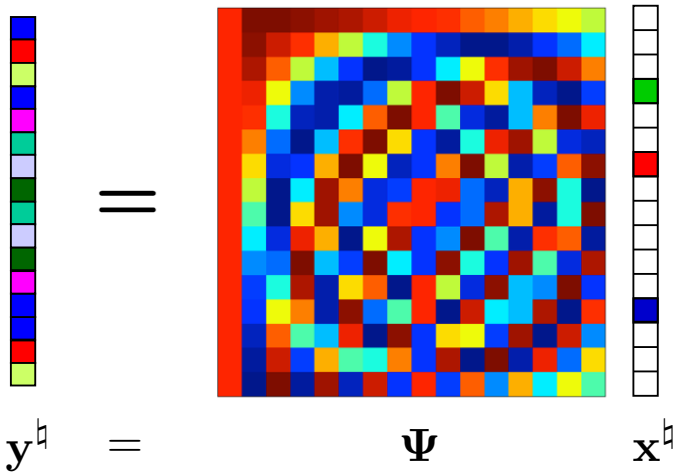
- The statistical error $\left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2$ can also be arbitrarily large when $\mathbf{w} \neq \mathbf{0}$. Hence, the solution is also not robust.

A natural signal model

Definition (s-sparse vector)

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

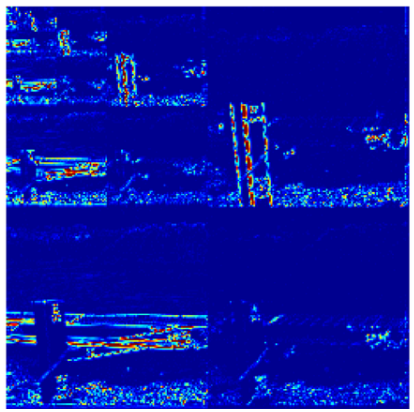




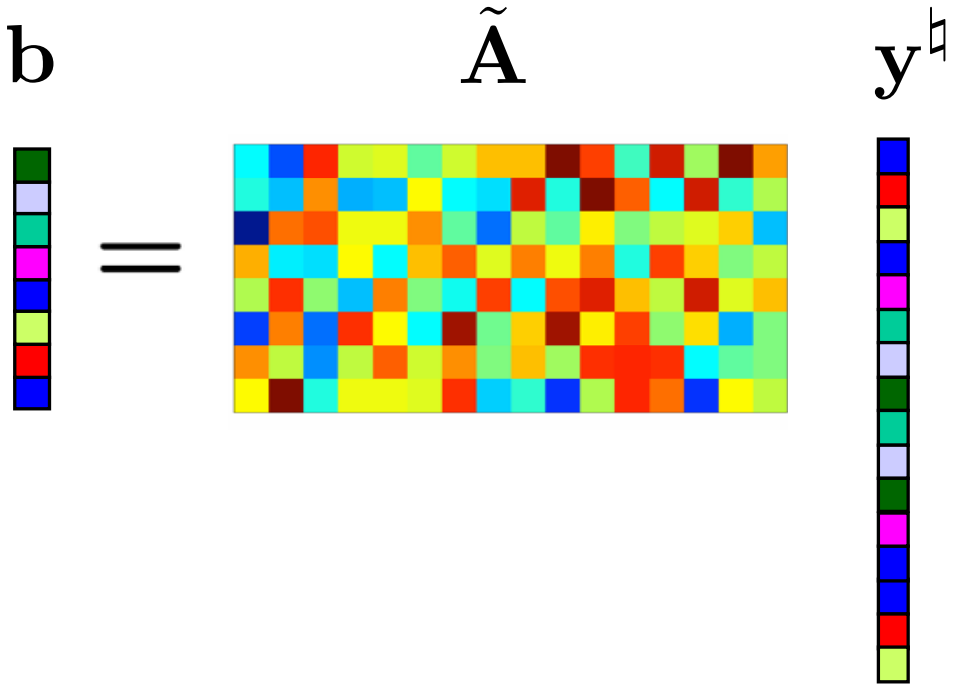
Sparse representations

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

- 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...

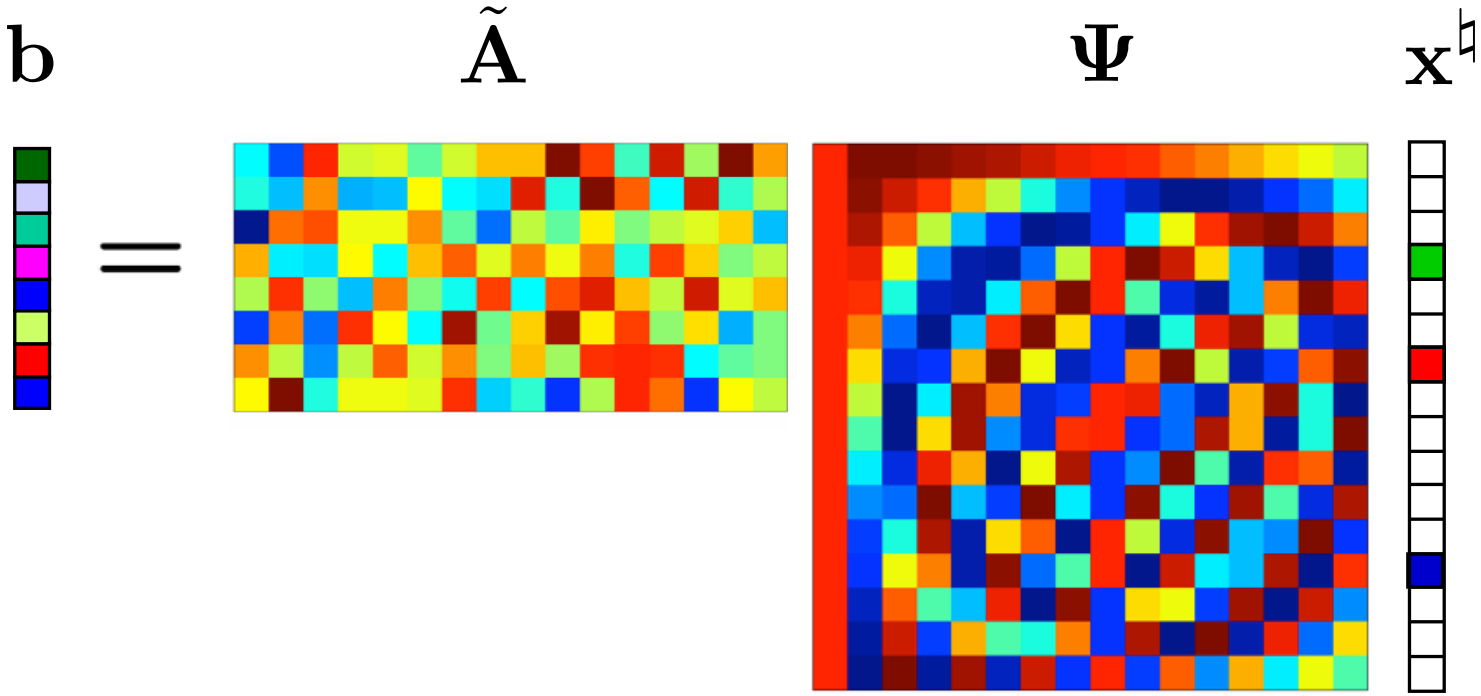


Sparse representations strike back!



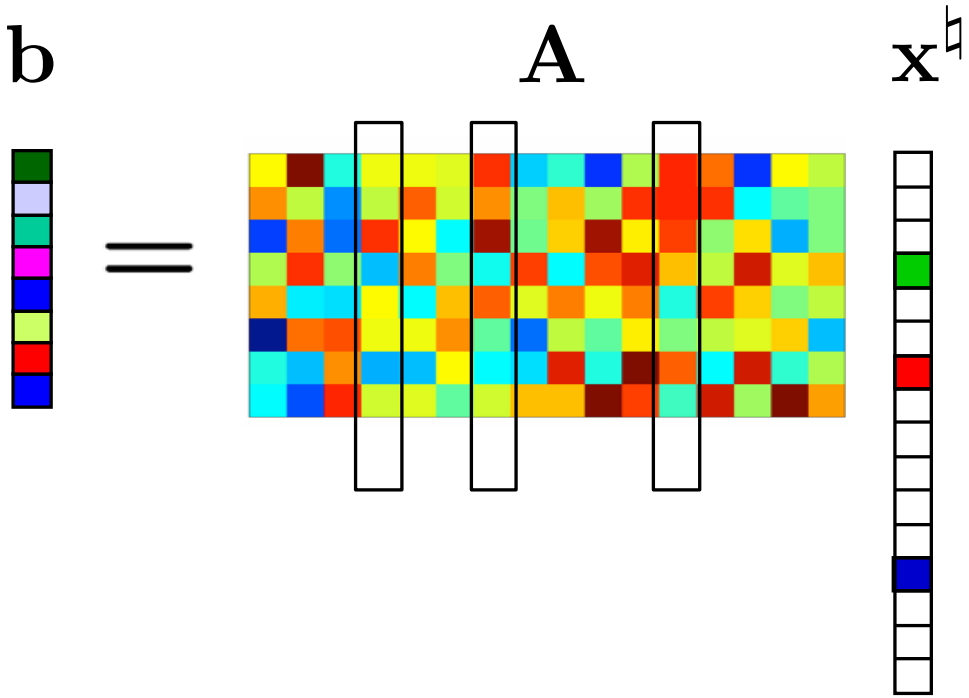
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



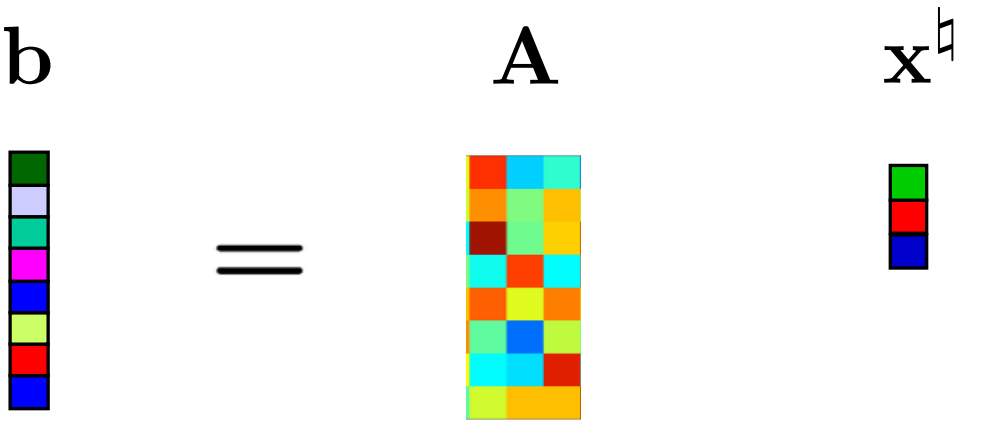
- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\Psi \in \mathbb{R}^{p \times p}$, $\mathbf{x}^\natural \in \mathbb{R}^p$, and $\|\mathbf{x}^\natural\|_0 \leq s < n$

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!



$n \times 1$ $\quad$ $n \times s$ $\quad$ $s \times 1$

A fundamental impact:

The matrix $\mathbf{A}$ effectively becomes *overcomplete*.

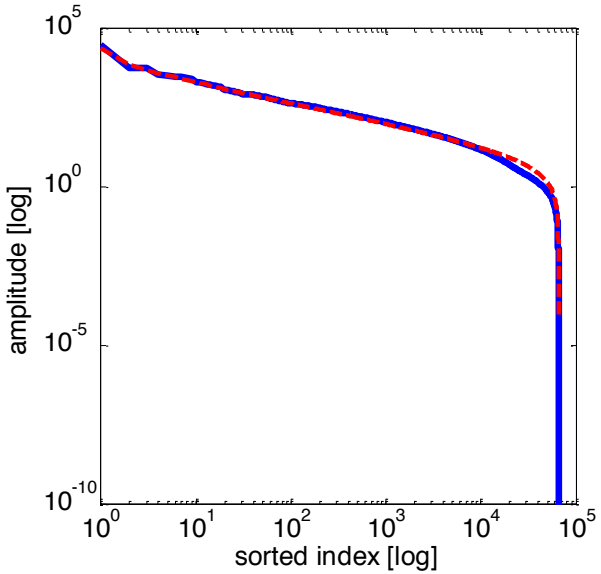
We could solve for $\mathbf{x}^\natural$ if we knew *the location of the non-zero entries of* $\mathbf{x}^\natural$.

Compressible signals

Real signals may not be exactly sparse, but approximately sparse, or *compressible*.

Roughly speaking, a vector $\mathbf{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is compressible if the number of its significant components, $|\{k : |x_k| \geq t, 1 \leq k \leq p\}|$, is small.

- **Cameraman**@MIT.



- **Solid curve**: Sorted wavelet coefficients of the cameraman image.
- **Dashed curve**: Expected order statistics of generalized Pareto distribution with shape parameter 1.67.

Compressible signals

Real signals may not be exactly sparse, but approximately sparse, or *compressible*.

Roughly speaking, a vector $\mathbf{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is compressible if the number of its significant components, $|\{k : |x_k| \geq t, 1 \leq k \leq p\}|$, is small.

Model: compressible signals tend to have small $w\ell_q$-quasi norms.

Definition (Weak ℓ_q-quasi norm)

$$\|\mathbf{x}\|_{w\ell_q} := \inf \left\{ M \geq 0 : |\{k : |x_k| \geq t, 1 \leq k \leq p\}| \leq \frac{M^q}{t^q} \text{ for all } t > 0 \right\}.$$

An equivalent definition

$$\|\mathbf{x}\|_{w\ell_q} = \max_{k \in \{1,\ldots,p\}} \left\{ k^{1/q} \, |x_k^*| \right\},$$

where $\left|x_k^*\right|$ denotes the k-th largest absolute value of the elements of $\mathbf{x}$.

Compressible signals contd.

Definition (Best s-term approximation error)

Let $\mathbf{x} \in \mathbb{R}^p$. For any $q > 0$, the best s-term approximation error is defined as follows

$$\sigma_s(\mathbf{x})_q := \inf_{\mathbf{z} \in \mathbb{R}^p} \left\{ \|\mathbf{x} - \mathbf{z}\|_q : \|\mathbf{z}\|_0 \leq s \right\}.$$

Proposition (Best s-term approximation error of signals in $w\ell_q$-space)

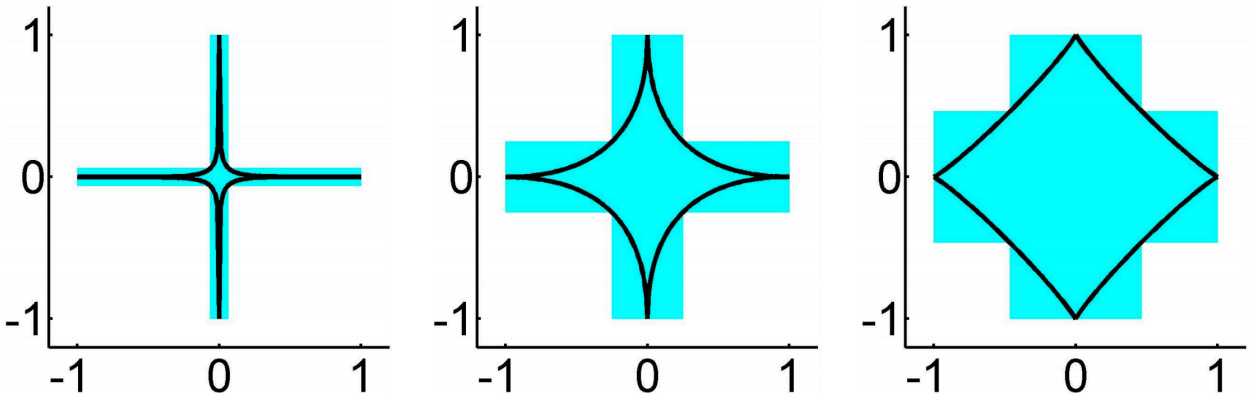
Let $\mathbf{x} \in \mathbb{R}^p$. For any $r > q > 0$, the

$$\sigma_s(\mathbf{x})_r \leq \frac{c_{q,r}}{s^{1/q-1/r}} \|\mathbf{x}\|_{w\ell_q},$$

where $c_{q,r} := \left[q(r-q)^{-1} \right]^{1/r}$.

- The proposition provides a justification for characterizing compressible signals based on weak ℓ_p-quasi norms.
- If $\mathbf{x} \in w\ell_q(R)$ lives in the $w\ell_q$-space with radius R (i.e., $\|\mathbf{x}\|_{w\ell_q} \leq R$), then we have $\left| x_k^* \right| \leq Rk^{-1/q}$ (i.e., its sorted coefficients exhibit a power law decay).

Example: the weak ℓ_q-quasi norm



Example $w\ell_q(1)$ balls: (Left) $q = 0.25$. (Middle) $q = 0.5$. (Right) $q = 0.9$.

Remark

Geometrically, the $w\ell_q(R)$-quasi norm ball includes the $\ell_q(R)$-quasi norm ball for all R:

$$w\ell_q(R) := \left\{ \mathbf{x} : \|\mathbf{x}\|_{w\ell_q} \leq R, \mathbf{x} \in \mathbb{R}^2 \right\},$$
$$\ell_q(R) := \left\{ \mathbf{x} : \|\mathbf{x}\|_{q} \leq R, \mathbf{x} \in \mathbb{R}^2 \right\}.$$

This observation also follows from the fact that $\|\mathbf{x}\|_{w\ell_q} \leq \|\mathbf{x}\|_q$ since it holds that $\|\mathbf{x}\|_q^q = \sum_{j=1}^{p} \left|x_j^*\right|^q \geq \sum_{j=1}^{k} \left|x_j^*\right|^q \geq k \left|x_k^*\right|^q = \|\mathbf{x}\|_{w\ell_q}^q$.

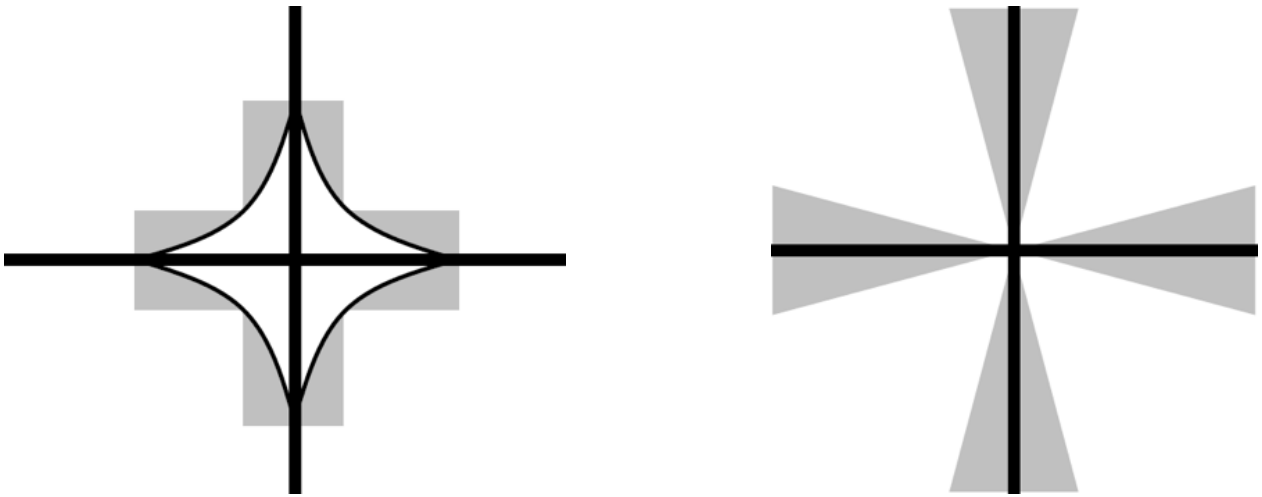
Compressibility via relative best s-term approximation error

A vector $\mathbf{x} \in \mathbb{R}^n$ compressible if it has small relative best s-term approximation error.

Definition (Relative best s-term approximation error [23])

Let $\mathbf{x} \in \mathbb{R}^n$. For any $p > 0$, the relative best s-term approximation error is defined as

$$\bar{\sigma}_s(\mathbf{x})_q := \frac{\sigma_s(\mathbf{x})_q}{\|\mathbf{x}\|_q}.$$



(Left) $\{\mathbf{x} : \|\mathbf{x}\|_{w\ell_q} \leq R\}$. (Right) $\{\mathbf{x} : \bar{\sigma}_s(\mathbf{x})_q \leq \varepsilon\}$

- The relative approximation model is arguably a better representation of compressibility.

Compressibility via relative best s-term approximation error

A vector $\mathbf{x} \in \mathbb{R}^n$ compressible if it has small relative best s-term approximation error.

Definition (Relative best s-term approximation error [23])

Let $\mathbf{x} \in \mathbb{R}^n$. For any $p > 0$, the relative best s-term approximation error is defined as

$$\bar{\sigma}_s(\mathbf{x})_q := \frac{\sigma_s(\mathbf{x})_q}{\|\mathbf{x}\|_q}.$$

Definition (Compressible distributions [23])

Let $\mathbf{x}$ such that its entries are i.i.d. from $x_i \sim P(x)$. The probability distribution function $P(x)$ is called q-compressible with parameters (ε, κ) if the following holds

$$\lim_{p \to \infty} \bar{\sigma}_{s_p}(\mathbf{x})_q \leq \varepsilon$$

almost surely for any sequence s_p such that $\lim_{p\to\infty} \frac{s_p}{p} \geq \kappa$, where $\varepsilon \ll 1$ and $\kappa \ll 1$.

- See [23] further for examples and illustrations of this concrete connection between deterministic sparsity models and probabilistic distributions.
- The nonsmooth convex formulations in this lecture are also useful in this context.

Example: Student's t distribution

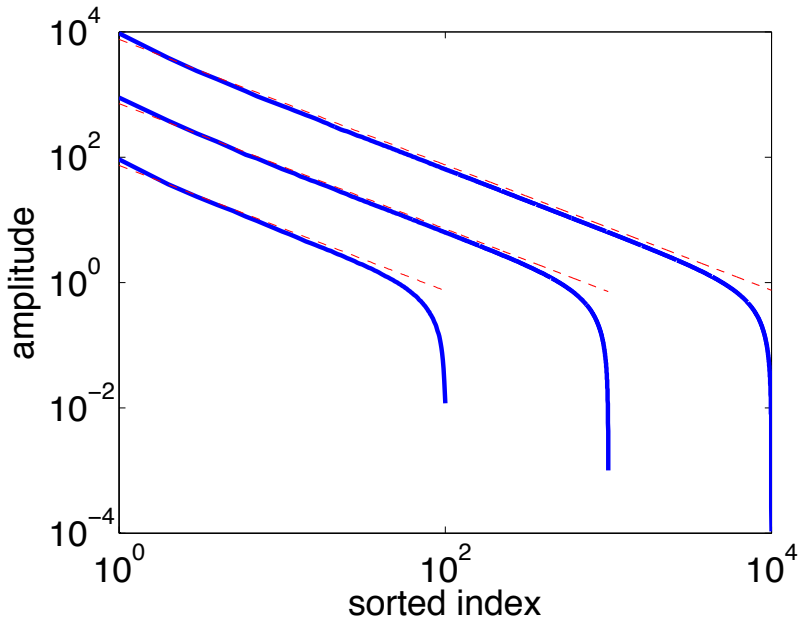
Definition (Compressible distributions [23])

Let $\mathbf{x}$ such that its entries are i.i.d. from $x_i \sim P(x)$. The probability distribution function $P(x)$ is called q-compressible with parameters (ε, κ) if the following holds

$$\lim_{p \to \infty} \bar{\sigma}_{s_p}(\mathbf{x})_q \leq \varepsilon$$

almost surely for any sequence s_p such that $\lim_{p\to\infty} \frac{s_p}{p} \geq \kappa$, where $\varepsilon \ll 1$ and $\kappa \ll 1$.

- Let $P(x) \propto (1 + x^2)^{-\frac{r+1}{2}}$ be the Student's t distribution.
- It is easy to verify that $P(x)$ is a compressible distribution.
- We also have $\mathbf{x} \in w\ell_q(R)$ where $q = r$ and $R \propto p^{1/q}$.
- Maximum a posteriori (MAP) estimation of compressible priors give rise to so-called *reweighted* methods.
- MAP with Student's t with the linear observation model leads to reweighted least squares algorithm.



Median of 100 realizations of $P(x)$ with $q = 1$.

A different tale of the linear model $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{w}$

A *realistic* linear model

Let $\mathbf{b} := \tilde{\mathbf{A}}\mathbf{y}^\natural + \tilde{\mathbf{w}} \in \mathbb{R}^n$.

- Let $\mathbf{y}^\natural := \Psi \mathbf{x}_{\text{real}} \in \mathbb{R}^m$ that admits a *compressible* representation $\mathbf{x}_{\text{real}}$.
- Let $\mathbf{x}_{\text{real}} \in \mathbb{R}^p$ that is *compressible* and let $\mathbf{x}^\natural$ be its *best s-term approximation*.
- Let $\tilde{\mathbf{w}} \in \mathbb{R}^n$ denote the possibly nonzero *noise* term.
- Assume that $\Psi \in \mathbb{R}^{m \times p}$ and $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times m}$ are known.

Then we have

$$\begin{aligned} \mathbf{b} &= \tilde{\mathbf{A}}\Psi\left(\mathbf{x}^\natural + \mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right) + \tilde{\mathbf{w}}. \\ &:= \underbrace{\left(\tilde{\mathbf{A}}\Psi\right)}_{\mathbf{A}}\mathbf{x}^\natural + \underbrace{\left[\tilde{\mathbf{w}} + \tilde{\mathbf{A}}\Psi\left(\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right)\right]}_{\mathbf{w}}, \end{aligned}$$

equivalently, $\boxed{\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}}$.

Peeling the onion

The *realistic* linear model uncovers yet another level of difficulty

Practical performance

The practical performance is determined by

$$\|\mathbf{x}_\epsilon^\star - \mathbf{x}_{\text{real}}\|_2 \leq \underbrace{\|\mathbf{x}_\epsilon^\star - \hat{\mathbf{x}}\|_2}_{\text{approximation error}} + \underbrace{\left\|\hat{\mathbf{x}} - \mathbf{x}^\natural\right\|_2}_{\text{statistical error}} + \underbrace{\left\|\mathbf{x}_{\text{real}} - \mathbf{x}^\natural\right\|_2}_{\text{model error}}.$$

- A great deal of research goes into learning representations that renders the model error negligible while still keeping statistical error low.

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

$n = 2s$ is sufficient for correctness of recovery (noiseless)

Let $\Sigma_s = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}$. If $\mathbf{w} = \mathbf{0}$ and $\Sigma_{2s} \cap \mathrm{null}(\mathbf{A}) = \emptyset$ (i.e., any matrix $\mathbf{A}$ with $\mathrm{rank}(\mathbf{A}) \geq 2s$), then $\mathcal{P}_0$ can perfectly recover any s-sparse $\mathbf{x}^\natural$ (i.e., $\hat{\mathbf{x}} = \mathbf{x}^\natural$).

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

$n = 2s$ is sufficient for correctness of recovery (noiseless)

Let $\Sigma_s = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}$. If $\mathbf{w} = \mathbf{0}$ and $\Sigma_{2s} \cap \mathrm{null}(\mathbf{A}) = \emptyset$ (i.e., any matrix $\mathbf{A}$ with $\mathrm{rank}(\mathbf{A}) \geq 2s$), then $\mathcal{P}_0$ can perfectly recover any s-sparse $\mathbf{x}^\natural$ (i.e., $\hat{\mathbf{x}} = \mathbf{x}^\natural$).

Minimum number of samples (noiseless)

$n = s + 1$ is necessary for correctness of $\mathcal{P}_0$ when $\mathbf{A}$ is "random" and $\mathbf{w} = \mathbf{0}$.

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

$n = 2s$ is sufficient for correctness of recovery (noiseless)

Let $\Sigma_s = \{\mathbf{x} : \|\mathbf{x}\|_0 \leq s\}$. If $\mathbf{w} = \mathbf{0}$ and $\Sigma_{2s} \cap \mathrm{null}(\mathbf{A}) = \emptyset$ (i.e., any matrix $\mathbf{A}$ with $\mathrm{rank}(\mathbf{A}) \geq 2s$), then $\mathcal{P}_0$ can perfectly recover any s-sparse $\mathbf{x}^\natural$ (i.e., $\hat{\mathbf{x}} = \mathbf{x}^\natural$).

Minimum number of samples (noiseless)

$n = s + 1$ is necessary for correctness of $\mathcal{P}_0$ when $\mathbf{A}$ is "random" and $\mathbf{w} = \mathbf{0}$.

Catch: $\mathcal{P}_0$ is *NP-hard* in general [19].

Solving for $\hat{\mathbf{x}}$ enables one to solve the exact cover by 3-sets, which is *NP-complete*.

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

Tricky question

Can we find a deterministic matrix $\mathbf{A}$ with $n = 2s$ so that $\mathcal{P}_0$ is polynomial time?

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

Tricky question

Can we find a deterministic matrix $\mathbf{A}$ with $n = 2s$ so that $\mathcal{P}_0$ is polynomial time?

Answer: Yes

We can use a partial Vandermonde matrix $\mathbf{A} = \mathbf{V}_p$ with $n = 2s$ where

$$\mathbf{V}_p = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \dots & \alpha_p \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 & \dots & \alpha_p^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_1^{n-1} & \alpha_2^{n-1} & \alpha_3^{n-1} & \dots & \alpha_p^{n-1} \end{bmatrix},$$

and $\alpha_l^k = p^{-0.5} \mathrm{e}^{-\frac{i2\pi kl}{p}}$, corresponding to the discrete Fourier transform when $n = p$.

Use Prony's method for your polynomial time recovery!

Estimating sparse parameters by ℓ_0-minimization

A possible 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,

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

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

Tricky question

Can we find a deterministic matrix $\mathbf{A}$ with $n = 2s$ so that $\mathcal{P}_0$ is polynomial time?

Answer: Yes

We can use a partial Vandermonde matrix $\mathbf{A} = \mathbf{V}_p$ with $n = 2s$ and then use Prony's method for polynomial time recovery.

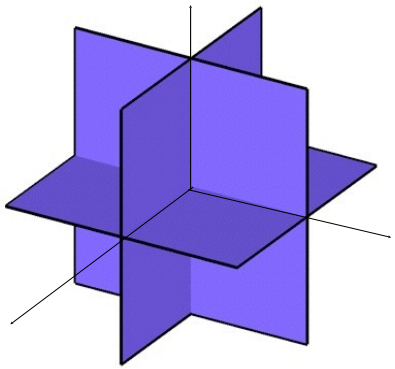
Catch: Instability of polynomial time solutions [19]

Recovery is not stable when $\mathbf{w} \neq \mathbf{0}$ and sensitive to any mismatched choices of s. Indeed, any stable recovery scheme requires $n = \Omega(s \log(ep/s))$.

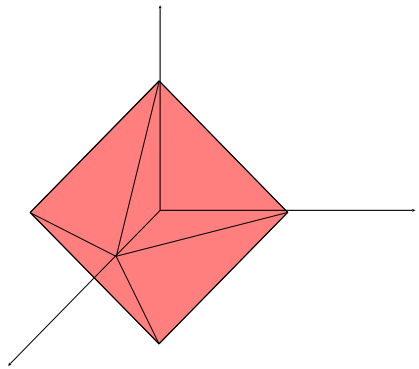
The ℓ_1-norm heuristic

Heuristic: *ℓ_1-ball with radius c_∞* is the "closest" convex set to the sparse vectors $\hat{\mathbf{x}} \in \left\{\mathbf{x} : \|\mathbf{x}\|_0 \leq s, \|\mathbf{x}\|_0 \leq c_\infty\right\}$ parameterized by their sparsity s and maximum amplitude c_∞.[1]

$$\hat{\mathbf{x}} \in \left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq c_\infty\right\} \quad \text{with some } c_\infty > 0.$$



The set
$\left\{\mathbf{x} : \|\mathbf{x}\|_0 \leq 2, \|\mathbf{x}\|_\infty \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$



The unit ℓ_1-norm ball
$\left\{\mathbf{x} : \|\mathbf{x}\|_1 \leq 1, \mathbf{x} \in \mathbb{R}^3\right\}$

This heuristic leads to the *basis pursuit denoising formulation*.

[1] We provide a mathematical interpretation of this heuristic via Lovász extension of set functions in Recitation 4.

Basis pursuit denoising (BPDN)

Definition (Basis pursuit denoising [15])

$$\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

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

Theorem (Existence of a stable solution in polynomial time [14])

This BPDN convex formulation is a second order cone program, which can be solved in polynomial time in terms of the inputs n and p (see Lecture 9). Surprisingly, if $\|\mathbf{w}\|_2 := \left\|\mathbf{b}-\mathbf{A}\mathbf{x}^\natural\right\|_2 \leq \kappa$, there exists an $\mathbf{A} \in \mathbb{R}^{n\times p}$ such that

$$\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}},$$

given that

$$n \geq \frac{2s\ln\left(\frac{p}{s}\right) + \frac{5}{4}s + \frac{3}{2}}{(1-\sqrt{\mu})^2},$$

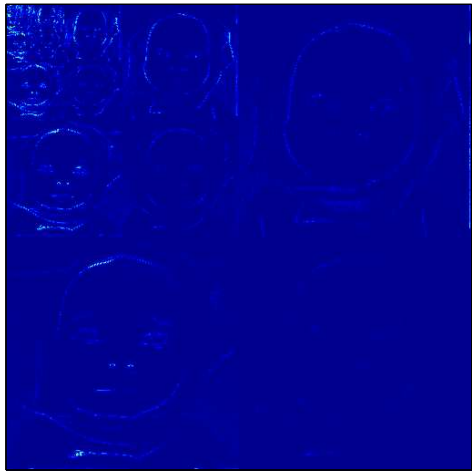
with some $\mu(\mathbf{A}) > 0$, which encodes the difficulty of the problem (more on this later).

Observation: It suffices to require $n = \Omega\left(s\log(p/s)\right)$.

Models with simplicity

p

pixels

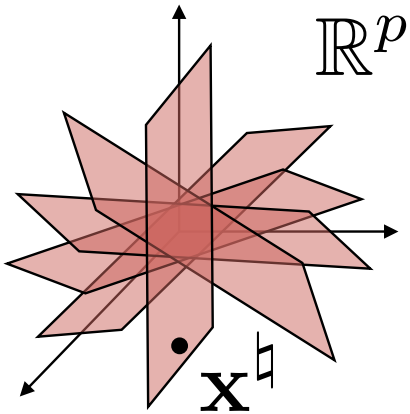


Information level:

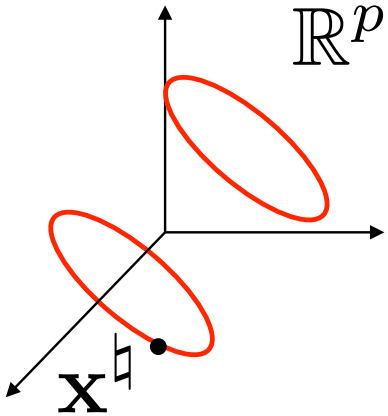
$s \ll p$

large wavelet coefficients

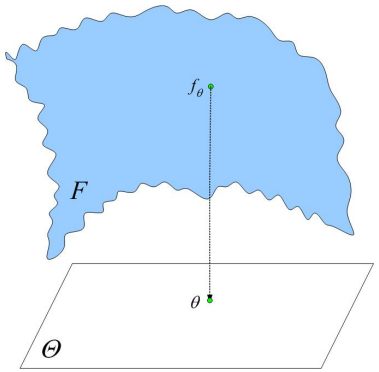
(blue = 0)



sparse signals



low-rank matrices



nonlinear models

Generalization via simple representations

Definition (Atomic sets & atoms)

An *atomic set* $\mathcal{A}$ is a set of vectors in $\mathbb{R}^p$. An *atom* is an element in an atomic set.

Terminology (Simple representation)

A parameter $\mathbf{x}^\natural \in \mathbb{R}^p$ *admits a simple representation with respect to an atomic set* $\mathcal{A} \subseteq \mathbb{R}^p$*, if it can be represented as a non-negative combination of few atoms, i.e.,* $\mathbf{x}^\natural = \sum_{i=1}^k c_i \mathbf{a}_i, \quad \mathbf{a}_i \in \mathcal{A},\ c_i \geq 0.$

Example (Sparse parameter)

Let $\mathbf{x}^\natural$ be s-sparse. Then $\mathbf{x}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, where $\mathbf{e}_i := (\delta_{1,i}, \delta_{2,i}, \ldots, \delta_{p,i})$ for all i.

Example (Sparse parameter with a dictionary)

Let $\Psi \in \mathbb{R}^{m \times p}$, and let $\mathbf{y}^\natural := \Psi \mathbf{x}^\natural$ be s-sparse. Then $\mathbf{y}^\natural$ can be represented as the non-negative combination of s elements in $\mathcal{A}$, with $\mathcal{A} := \{\pm\psi_1, \ldots, \pm\psi_p\}$, where ψ_k denotes the kth column of Ψ.

Simplest estimate

Recall the linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$.

To find the *simplest estimate* of $\mathbf{x}^\natural$ with respect to an atomic set $\mathcal{A}$ leads to the following formulation.

Possible approach

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ k : \mathbf{x} = \sum_{i=1}^{k} c_i \mathbf{a}_i,\ c_i \geq 0, \mathbf{a}_i \in \mathcal{A}, \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}.$$

However, when $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, we have an equivalent formulation

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\},$$

which is *NP-hard*.

Atomic norm

Recall how we get around the NP-hardness issue.

Definition (Basis pursuit denoising [15])

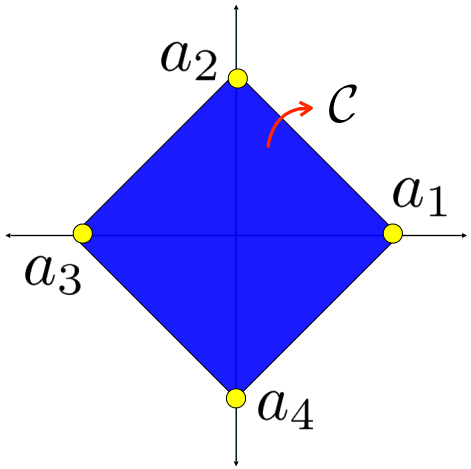
$$\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

with some $\kappa \geq 0$.

We observe that the ℓ_1-norm is the *atomic norm* associated with the atomic set $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, which is indeed the convex hull of the set.

$$\mathcal{A} := \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}.$$

$$\mathcal{C} := \operatorname{conv}(\mathcal{A}).$$



Gauge functions and atomic norms

Definition (Gauge function)

Let $\mathcal{C}$ be a convex set in $\mathbb{R}^p$, the **gauge function** associated with $\mathcal{C}$ is given by

$$g_{\mathcal{C}}(\mathbf{x}) := \arg\inf_{t>0} \{\mathbf{x} = t\mathbf{c} : \mathbf{c} \in \mathcal{C}, \quad \forall \mathbf{x} \in \mathbb{R}^p\}.$$

Definition (Atomic norm)

Let $\mathcal{A}$ be a symmetric *atomic set* in $\mathbb{R}^p$ such that if $\mathbf{a} \in \mathcal{A}$ then $-\mathbf{a} \in \mathcal{A}$ for all $\mathbf{a} \in \mathcal{A}$. Then, the **atomic norm** associated with a symmetric atomic set $\mathcal{A}$ is given by

$$\|\mathbf{x}\|_{\mathcal{A}} := g_{\mathrm{conv}(\mathcal{A})}(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^p,$$

where $\mathrm{conv}(\mathcal{A})$ denotes the *convex hull* of $\mathcal{A}$.

Example

1. Let $\mathcal{A}$ be the set of unit-normed one-sparse vectors, then $\|\mathbf{x}\|_{\mathcal{A}} = \|\mathbf{x}\|_1$.
2. Let $\mathcal{A} = \{\pm 1\}_{i=1}^p$, then $\|\mathbf{x}\|_{\mathcal{A}} = \|\mathbf{x}\|_\infty$.

*Atomic norms and gauge functions contd.

- Gauge functions are not norms in general unless the inducing atomic set satisfies the **centrally symmetric** condition:

$$\mathbf{x} \in \mathcal{A} \text{ if and only if } -\mathbf{x} \in \mathcal{A}$$

Example

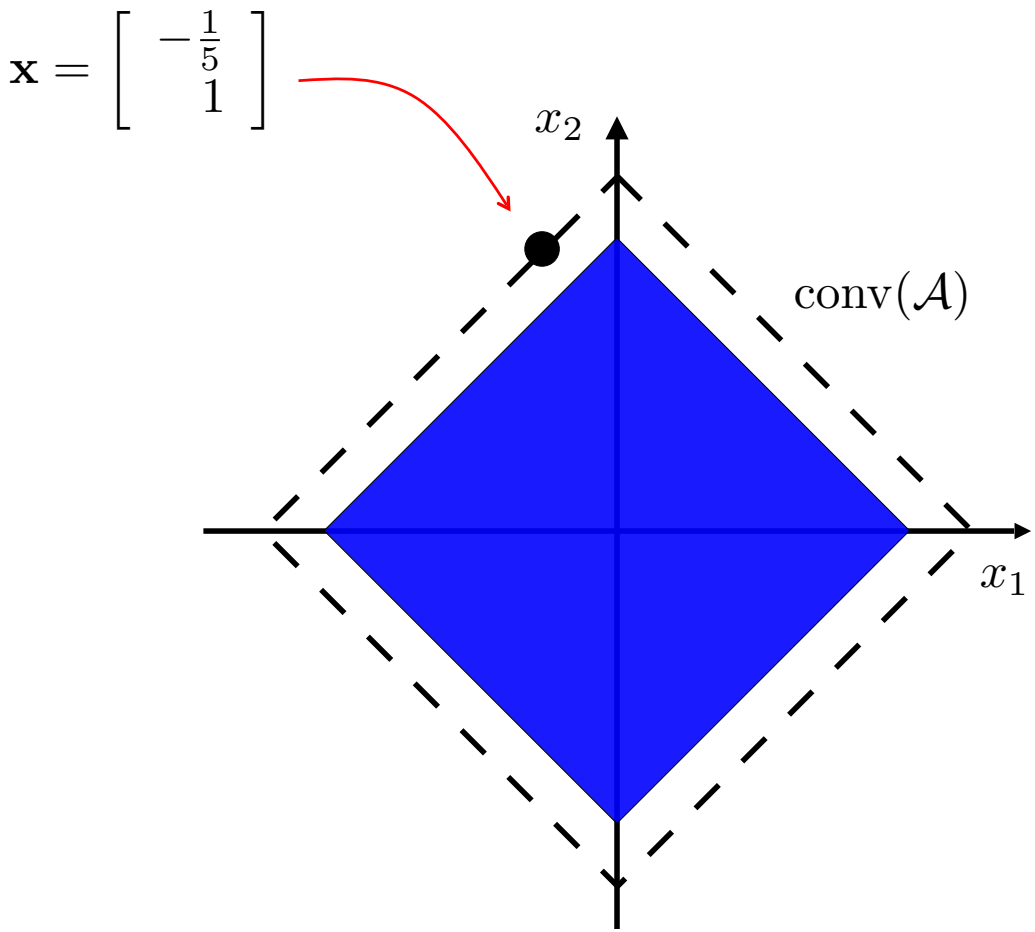
1. Let $\mathcal{A} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$. Then $\|\cdot\|_{\mathcal{A}} = g_{\mathrm{conv}(\mathcal{A})}(\cdot)$ is a norm.
2. Let $\mathcal{A} = \left\{ \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}$. Then $\|\cdot\|_{\mathcal{A}} = g_{\mathrm{conv}(\mathcal{A})}(\cdot)$ is *not* a norm.

Proposition

A gauge function associated with a non-empty atomic set $\mathcal{A}$ is a norm if and only if $\mathcal{A}$ is centrally symmetric.

Pop quiz 1

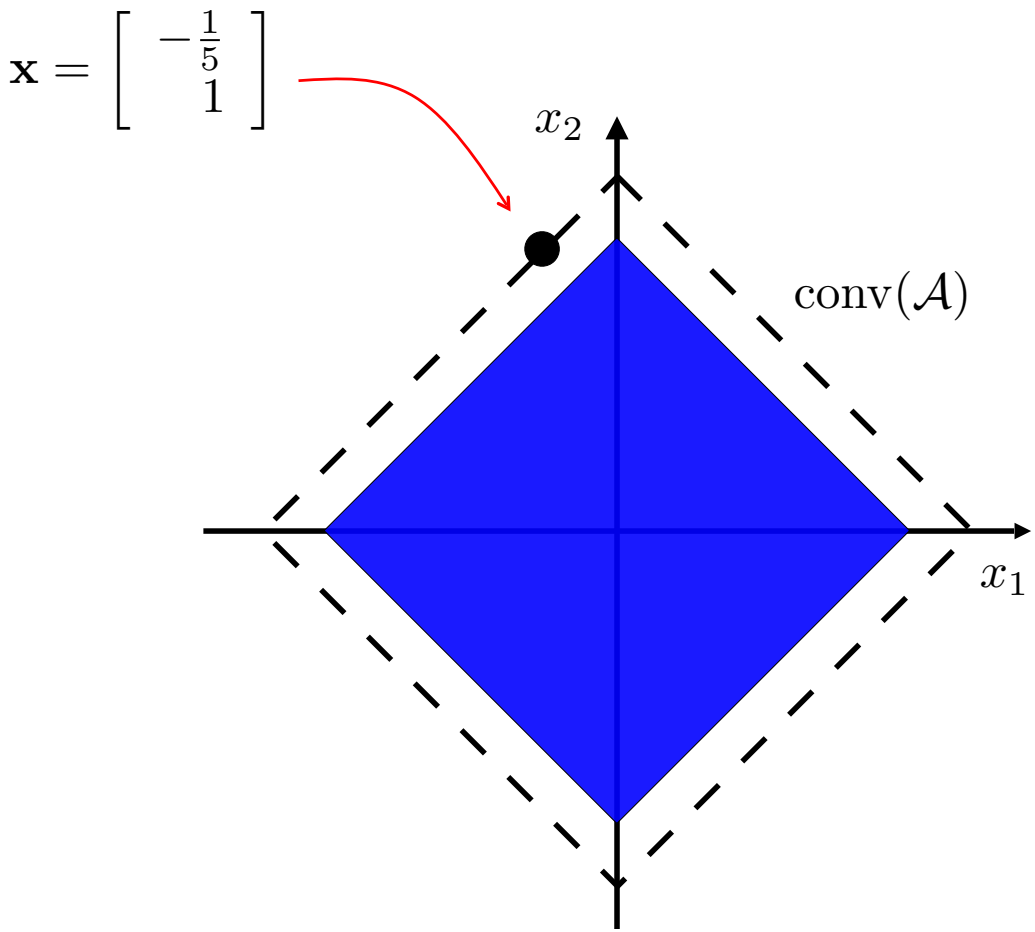
Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\mathbf{x}\|_{\mathcal{A}}$?



Pop quiz 1

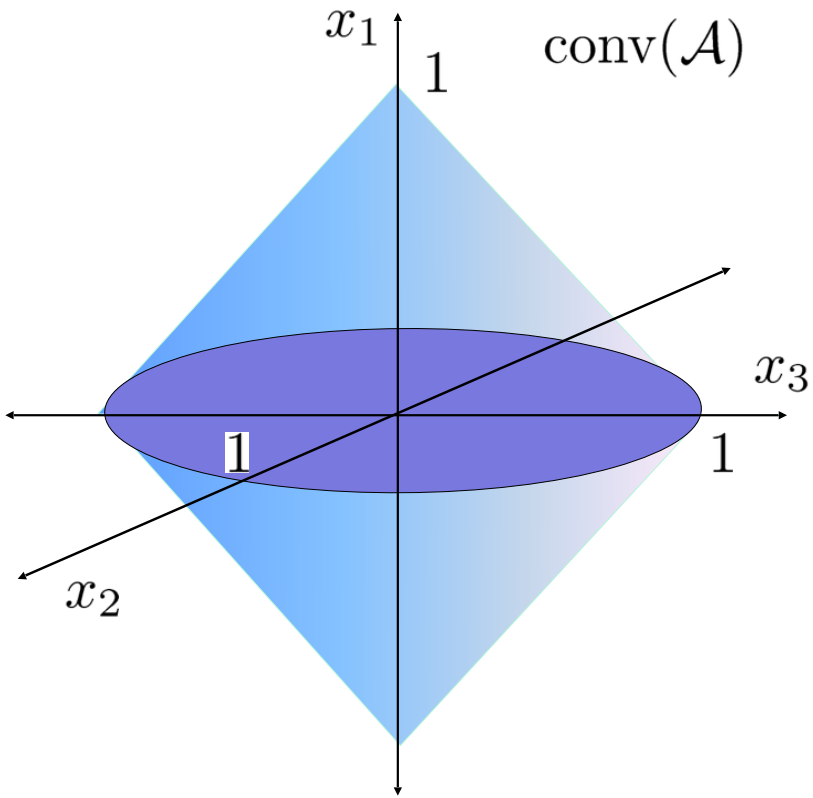
Let $\mathcal{A} := \left\{(1,0)^T, (0,1)^T, (-1,0)^T, (0,-1)^T\right\}$, and let $\mathbf{x} := (-\frac{1}{5}, 1)^T$. What is $\|\mathbf{x}\|_{\mathcal{A}}$?

ANS: $\|\mathbf{x}\|_{\mathcal{A}} = \frac{6}{5}$.



Pop quiz 2

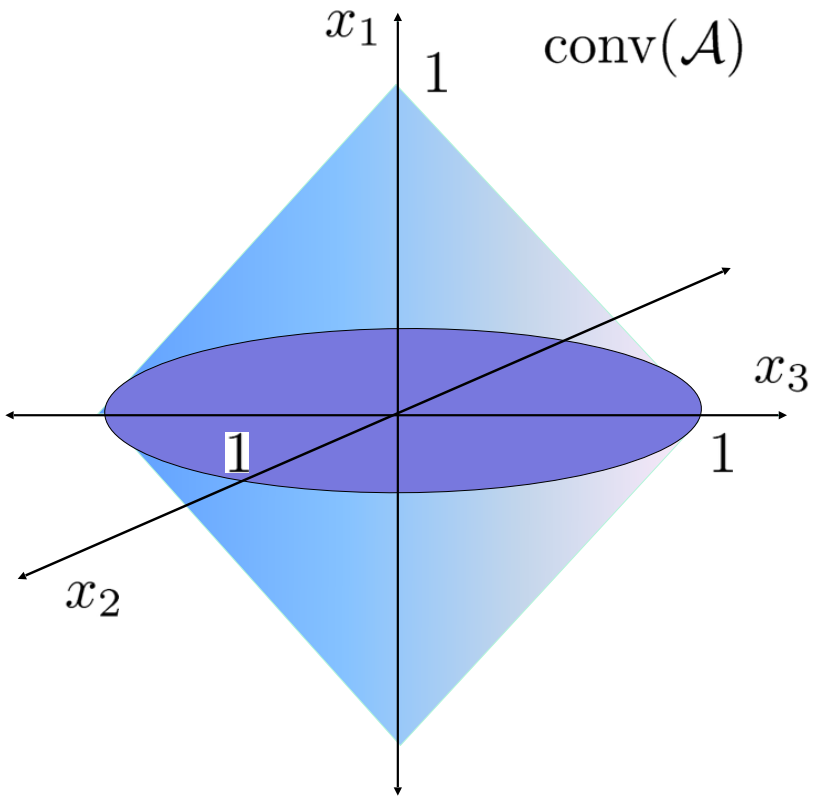
What is the expression of $\|\mathbf{x}\|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?



Pop quiz 2

What is the expression of $\|\mathbf{x}\|_{\mathcal{A}}$ for any $\mathbf{x} := (x_1, x_2, x_3)^T \in \mathbb{R}^3$?

ANS: $\|\mathbf{x}\|_{\mathcal{A}} = |x_1| + \left\|(x_2, x_3)^T\right\|_2$.



Basis pursuit with atomic norms

Linear model with *simple* parameter

Let $\mathcal{A}$ be an atomic set in $\mathbb{R}^p$. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ be *simple* with respect to $\mathcal{A}$, and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. The samples are given by $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes the unknown noise.

We consider the following estimator.

Basis pursuit denoising with atomic norms

$$\hat{\mathbf{x}}_{\text{BPDN}} := \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

with some $\kappa \geq 0$.

- In general, this problem cannot be solved in polynomial time even if it is convex (see Recitation 1).
- When we can solve it, this heuristic formulation provides surprisingly good results.

Performance guarantee of basis pursuit denoising

Theorem

Recall

$$\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

If $\|\mathbf{w}\|_2 := \left\|\mathbf{b}-\mathbf{A}\mathbf{x}^\natural\right\|_2 \leq \kappa$, *it is possible to have*

$$\left\|\hat{\mathbf{x}}_{BPDN}-\mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}},$$

given that

$$n \geq \frac{w^2+\frac{3}{2}}{\left(1-\sqrt{\mu}\right)^2},$$

with some $\mu(\mathbf{A})>0$, *where* w *is some function of the atomic set* $\mathcal{A}$ *and* $\mathbf{x}^\natural$.

- The quantity w^2 characterizes the *degrees-of-freedom* of $\mathbf{x}^\natural$.
- The parameter $\mu(\mathbf{A})$ characterizes the *well-posedness* of the estimation problem.

We prove the theorem in the following slides.

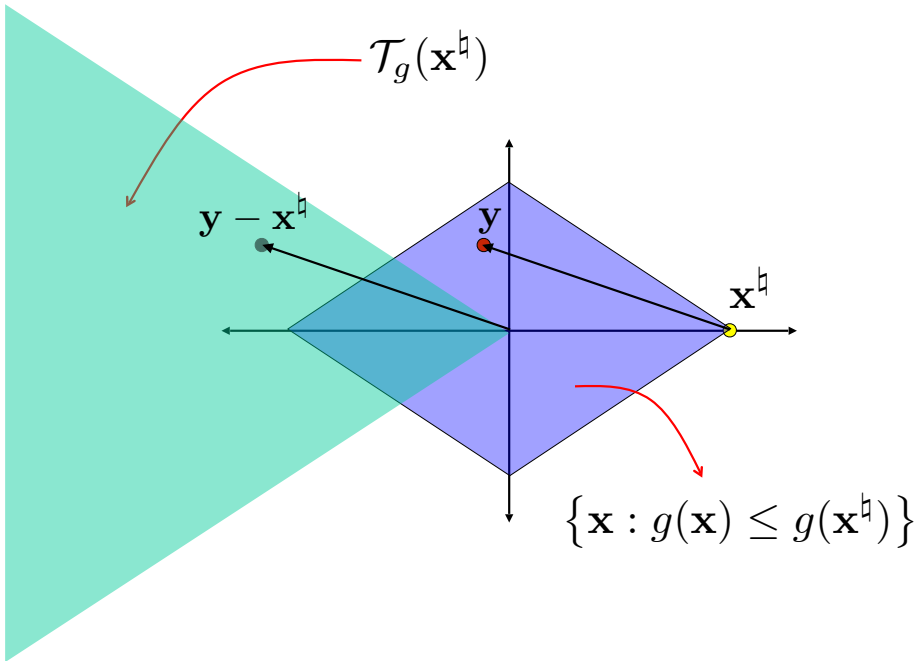
First we need the notion of *tangent cones*.

Tangent cone

Definition (Tangent cone)

Let $g : \mathbb{R}^p \to \mathbb{R} \cup \{-\infty, +\infty\}$ be a proper lower semi-continuous convex function. The tangent cone $\mathcal{T}_g(\mathbf{x})$ of the function g at a point $\mathbf{x} \in \mathbb{R}^p$ is defined as

$$\mathcal{T}_g(\mathbf{x}) := \operatorname{cone}\{\mathbf{y} - \mathbf{x} : g(\mathbf{y}) \leq g(\mathbf{x}), \mathbf{y} \in \mathbb{R}^p\}.$$



Condition for exact recovery in the *noiseless* case

We consider estimating $\mathbf{x}^\natural \in \mathbb{R}^p$, which is simple with respect to an atomic set $\mathcal{A}$, given samples $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ and $\mathbf{A} \in \mathbb{R}^{n \times p}$, $n \leq p$, by

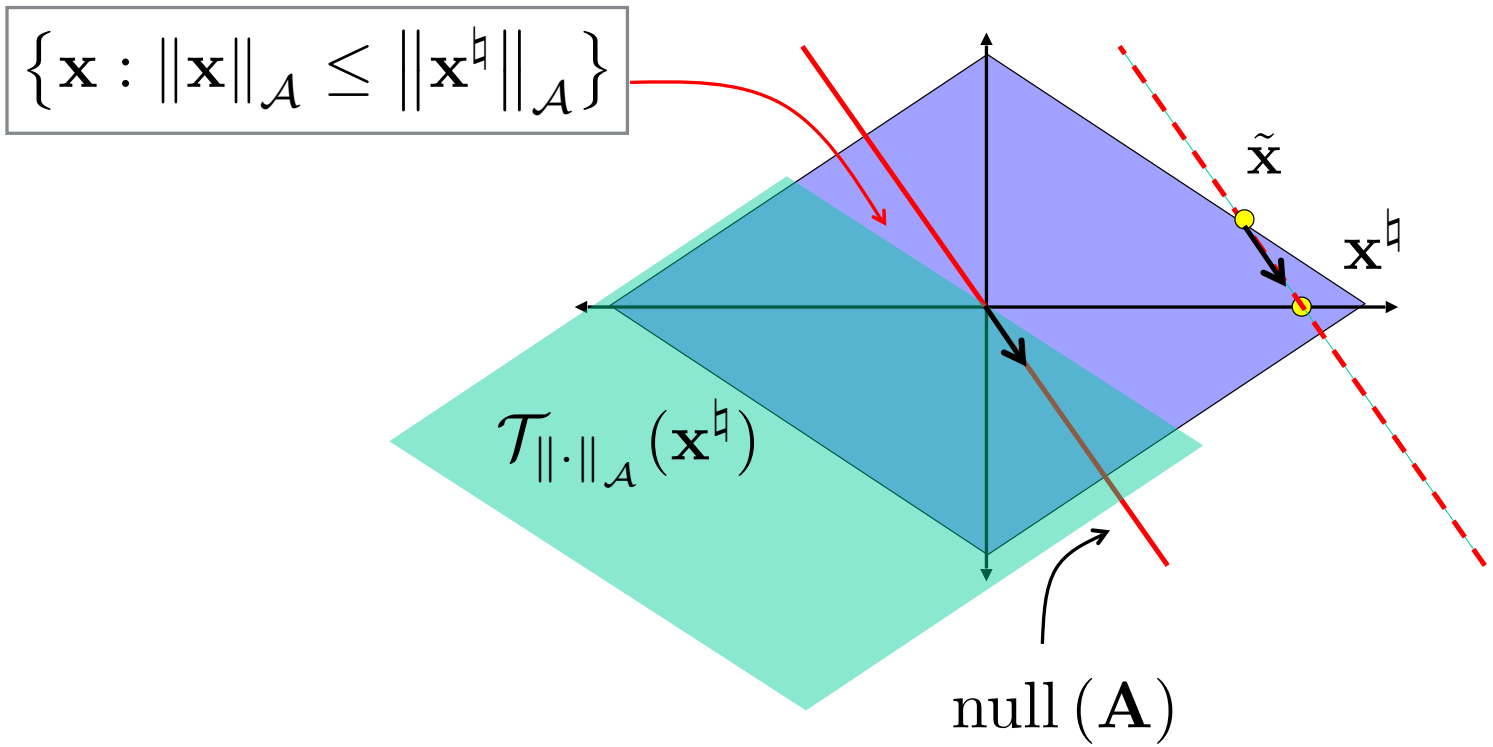
$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}.$$

Condition for exact recovery in the *noiseless* case

Proposition

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *Recall* $\hat{\mathbf{x}}_{BPDN} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}$.

We have $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ *if and only if* $\mathcal{T}_g\left(\mathbf{x}^\natural\right) \cap \operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$.

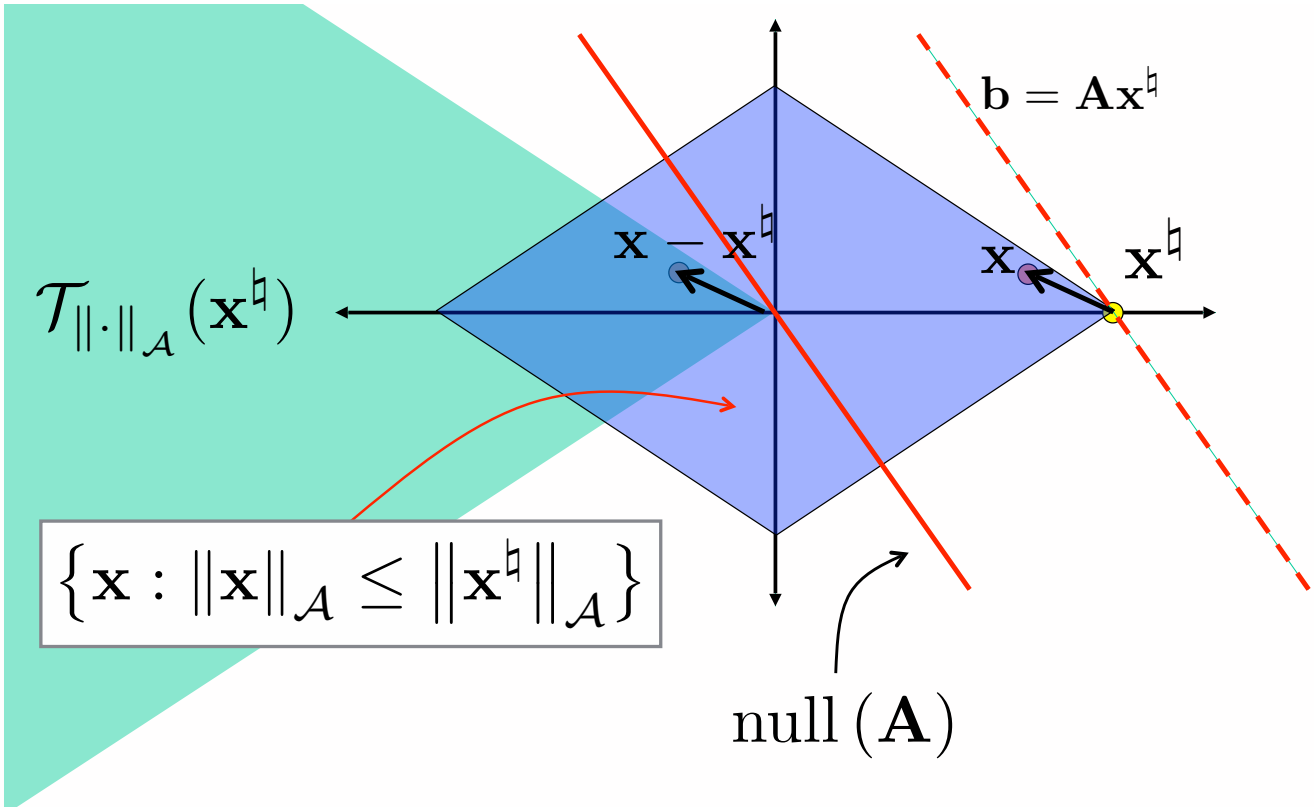


Condition for exact recovery in the *noiseless* case

Proposition

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *Recall* $\hat{\mathbf{x}}_{BPDN} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}$.

We have $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ *if and only if* $\mathcal{T}_g\left(\mathbf{x}^\natural\right) \cap \operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$.



Condition for exact recovery in the *noisy* case

We consider estimating $\mathbf{x}^\natural \in \mathbb{R}^p$, which is simple with respect to an atomic set $\mathcal{A}$, given samples $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ and $\mathbf{A} \in \mathbb{R}^{n \times p}$, $n \leq p$, where $\mathbf{w}$ denotes the unknown noise, by

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}.$$

Condition for good recovery in the *noisy* case

Definition (Restricted strong convexity)

The restricted strong convexity condition holds if $\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.

Proposition

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *Recall* $\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$.

We have $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}}$ *if* $\|\mathbf{w}\|_2 \leq \kappa$ *and the restricted strong convexity condition holds with some* $\mu > 0$.

Condition for good recovery in the *noisy* case

Definition (Restricted strong convexity)

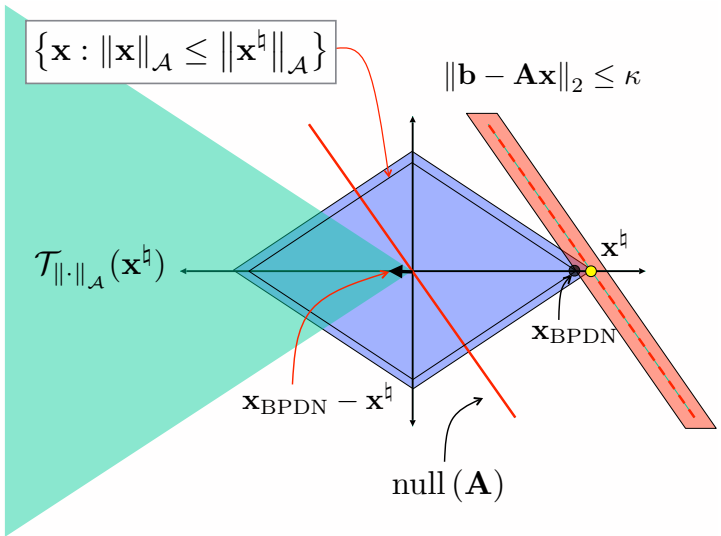
The restricted strong convexity condition holds if $\|\mathbf{Az}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.

Proposition

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *Recall* $\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b} - \mathbf{Ax}\|_2 \leq \kappa \right\}$.

We have $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}}$ *if* $\|\mathbf{w}\|_2 \leq \kappa$ *and the restricted strong convexity condition holds with some* $\mu > 0$.

Trivial observation: $\hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$



Condition for good recovery in the *noisy* case

Definition (Restricted strong convexity)

The restricted strong convexity condition holds if $\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.

Proposition

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *Recall* $\hat{\mathbf{x}}_{BPDN} := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$.

We have $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}}$ *if* $\|\mathbf{w}\|_2 \leq \kappa$ *and the restricted strong convexity condition holds with some* $\mu > 0$.

Proof.

By definition $\hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$; thus

$$\left\|\mathbf{A}\left(\hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural\right)\right\|_2 \geq \sqrt{\mu}\left\|\hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural\right\|_2.$$

By the triangle inequality,

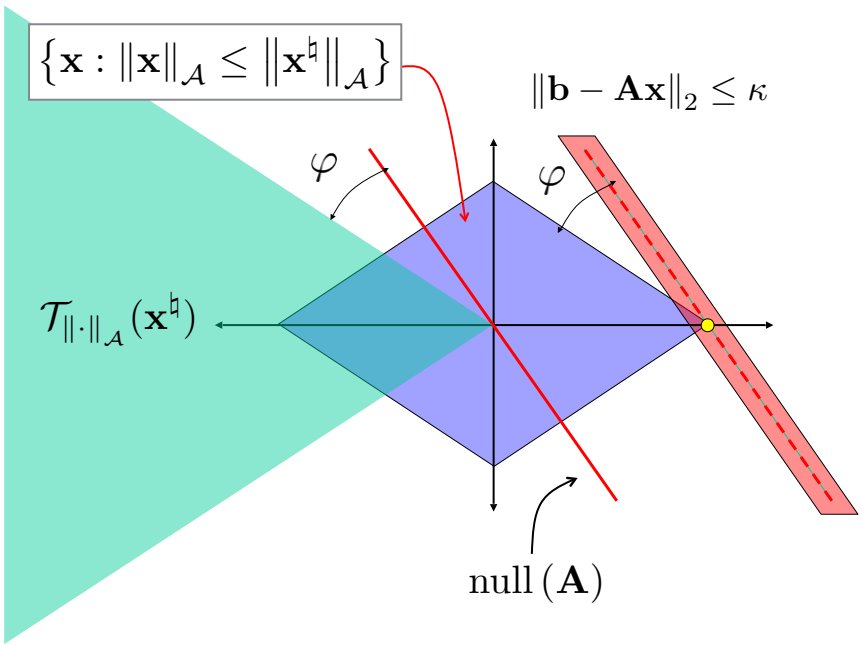
$$\left\|\mathbf{A}\left(\hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural\right)\right\|_2 \leq \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{BPDN}}\|_2 + \left\|\mathbf{b} - \mathbf{A}\mathbf{x}^\natural\right\|_2 \leq 2\kappa.$$

□

Condition for good recovery in the *noisy* case

Definition (Restricted strong convexity)

The restricted strong convexity condition holds if $\|\mathbf{Az}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.



- In the figure, μ is proportional to $\sin^2(\varphi)$, where the proportionality depends on the norm of the rows of $\mathbf{A}$.

Interpretation of the *restricted strong convexity* condition

Definition (Restricted strong convexity)

The restricted strong convexity condition holds if $\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.

Proposition

The restricted strong convexity condition holds if and only if the function $f : \mathbf{h} \mapsto \frac{1}{2}\left\|\mathbf{b} - \mathbf{A}\left(\mathbf{x}^\natural + \mathbf{h}\right)\right\|_2^2$ *satisfies*

$$f(\mathbf{x}^\natural + \mathbf{h}) \geq f(\mathbf{x}^\natural) + \left\langle \nabla f(\mathbf{x}^\natural), \mathbf{h} \right\rangle + \frac{\mu}{2}\|\mathbf{h}\|_2^2, \quad \text{for all } \mathbf{h} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right),$$

or, $f(\mathbf{h})$ *behaves as a strongly convex function for* $\mathbf{h} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$.

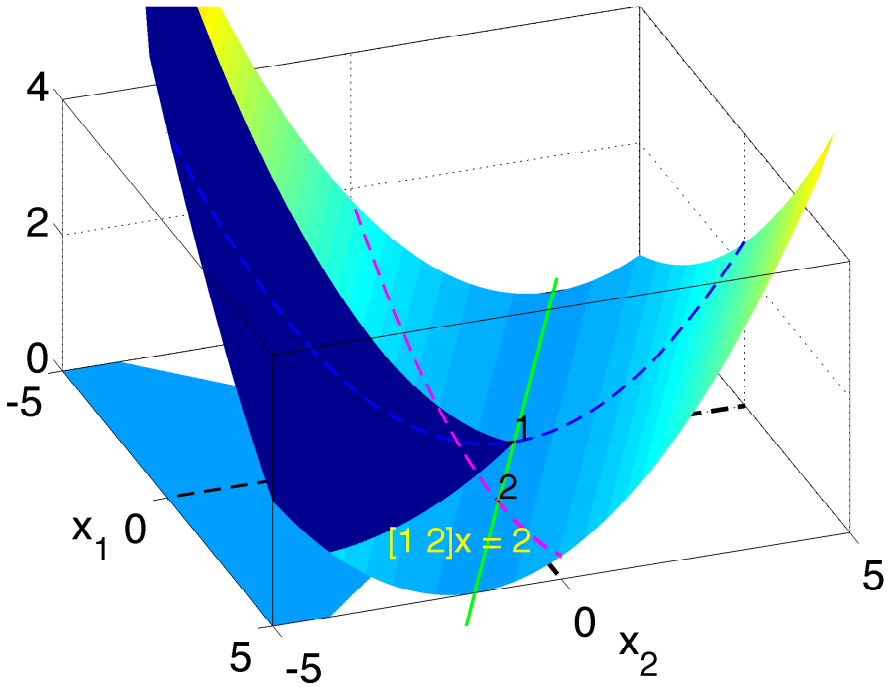
Interpretation of the *restricted strong convexity* condition

Proposition

The restricted strong convexity condition holds if and only if the function $f : \mathbf{h} \mapsto \frac{1}{2} \left\| \mathbf{b} - \mathbf{A} \left(\mathbf{x}^\natural + \mathbf{h} \right) \right\|_2^2$ *satisfies*

$$f(\mathbf{x}^\natural + \mathbf{h}) \geq f(\mathbf{x}^\natural) + \left\langle \nabla f(\mathbf{x}^\natural), \mathbf{h} \right\rangle + \frac{\mu}{2} \|\mathbf{h}\|_2^2 , \quad \text{for all } \mathbf{h} \in \mathcal{T}_g \left(\mathbf{x}^\natural \right),$$

or, $f(\mathbf{h})$ *behaves as a strongly convex function for* $\mathbf{h} \in \mathcal{T}_g \left(\mathbf{x}^\natural \right)$.



Interpretation of the *restricted strong convexity* condition

Definition (Restricted strong convexity)

The restricted strong convexity condition holds if $\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ for all $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ with some $\mu > 0$.

Proposition

The restricted strong convexity condition holds if and only if the function $f : \mathbf{h} \mapsto \frac{1}{2}\left\|\mathbf{b} - \mathbf{A}\left(\mathbf{x}^\natural + \mathbf{h}\right)\right\|_2^2$ *satisfies*

$$f(\mathbf{x}^\natural + \mathbf{h}) \geq f(\mathbf{x}^\natural) + \left\langle \nabla f(\mathbf{x}^\natural), \mathbf{h} \right\rangle + \frac{\mu}{2}\|\mathbf{h}\|_2^2, \quad \text{for all } \mathbf{h} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right),$$

or, $f(\mathbf{h})$ *behaves as a strongly convex function for* $\mathbf{h} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$.

Observation: Note that $\hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural + \mathbf{h}$ with some $\mathbf{h} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ by definition. Thus the restricted strong convexity condition implies that the function $\frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$ behaves as if $\mathbf{A}$ had full column rank for all possible values of $\hat{\mathbf{x}}_{\text{BPDN}}$.

- There are some variants of this restricted strong convexity condition based on similar ideas [4, 27].

Verifying the conditions

Now we have performance guarantees for $\hat{\mathbf{x}}_{\text{BPDN}}$.

Proposition (Noiseless)

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *We have* $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ *if and only if* $\mathcal{T}_g\left(\mathbf{x}^\natural\right) \cap \operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$.

Proposition (Noisy)

Let $g : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. *We have* $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}}$ *if* $\|\mathbf{w}\|_2 \leq \kappa$ *and* $\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2$ *for all* $\mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right)$ *with some* $\mu > 0$.

How do we verify *these conditions*, especially when we do not know $\mathbf{x}^\natural$ and thus $\mathcal{T}_g\left(\mathbf{x}^\natural\right)$?

No good answers currently.

The probabilistic approach

Show that no matter what $\mathbf{x}^\natural$ is, under *some other verifiable conditions*, we have

$$\mathcal{T}_g\left(\mathbf{x}^\natural\right) \cap \operatorname{null}\left(\mathbf{A}\right) = \{\mathbf{0}\}\,, \text{ or}$$
$$\|\mathbf{A}\mathbf{z}\|_2^2 \geq \mu \|\mathbf{z}\|_2^2\,, \quad \forall \mathbf{z} \in \mathcal{T}_g\left(\mathbf{x}^\natural\right) \text{ with some } \mu > 0,$$

with probability bounded away from 0.

The key technical tool is the *escape-through-the-mesh theorem*.

Escape-through-the-mesh theorem

Theorem (Escape-through-the-mesh theorem [14, 22, 34])

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ *be a matrix of i.i.d. Gaussian random variables with zero means and variances* $1/n$*. Let* Ω *be a given set on the unit* ℓ_2*-norm sphere. Then*

$$\mathbb{P}\left(\left\{\|\mathbf{A}\mathbf{x}\|_2 \geq \sqrt{\mu},\, \forall \mathbf{x} \in \Omega\right\}\right) \geq 1 - \exp\left\{-\frac{1}{2}\left[a_n - w(\Omega) - \sqrt{n\mu}\right]^2\right\}$$

given that $a_n - w(\Omega) - \sqrt{n\mu} \geq 0$*, where* $a_n := \sqrt{2}\,\Gamma\left(\frac{n+1}{2}\right)\Big/\Gamma\left(\frac{n}{2}\right)$*,* Γ *being the gamma function, and*

$$w(\Omega) := \mathbb{E}\left[\max_{\mathbf{x}} \{\langle \mathbf{g}, \mathbf{x}\rangle\} : \mathbf{x} \in \Omega\right],$$

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

Observation:

- The event $\left\{\|\mathbf{A}\mathbf{x}\|_2^2 \geq \mu,\, \forall \mathbf{x} \in \Omega\right\}$ implies the event that $\mathrm{null}\,(\mathbf{A})$ does not intersect with the mesh Ω.
- One can prove that $\frac{n}{\sqrt{n+1}} \leq a_n \leq \sqrt{n}$, which implies $a_n \approx \sqrt{n}$.

Proof of the escape-through-the-mesh theorem

First we note that $\left\{\|\mathbf{Ax}\|_2 \geq \sqrt{\mu},\, \forall \mathbf{x} \in \Omega\right\} = \left\{\min_{\mathbf{x}\in\Omega}\left\{\|\mathbf{Ax}\|_2\right\} \geq \sqrt{\mu}\right\}$ and $f : \mathbf{A} \in \mathbb{R}^{n\times p} \mapsto \min_{\mathbf{x}\in\Omega}\left\{\|\mathbf{Ax}\|_2\right\}$ is a Lipschitz function.

Proposition

Let $f : \mathbf{A} \in \mathbb{R}^{n\times p} \mapsto \min_{\mathbf{x}\in\Omega}\left\{\|\mathbf{Ax}\|_2\right\}$. *For all* $\mathbf{H} \in \mathbb{R}^{n\times p}$,

$$|f(\mathbf{A}+\mathbf{H}) - f(\mathbf{A})| \leq \|\mathbf{H}\|_F .$$

Thus f can be viewed as a Lipschitz function of np i.i.d. Gaussian random variables.

Theorem (Tsirelson-Ibragimov-Sudakov [6])

Let $\mathbf{g} \in \mathbb{R}^p$ *be a vector of i.i.d. standard Gaussian random variables. Let* $h : \mathbb{R}^p \to \mathbb{R}$ *be* K*-Lipschitz. Then for all* $t > 0$,

$$\mathbb{P}\left(\{h(\mathbf{g}) \leq \mathbb{E}\left[h(\mathbf{g})\right] - t\}\right) \leq \exp\left(-\frac{t^2}{2K}\right).$$

Proof of the escape-through-the-mesh theorem

The issue now is to evaluate $\mathbb{E}\left[f(\mathbf{A})\right]$.

Theorem (Gordon [22])

Let $\mathbf{G} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. standard Gaussian random variables with $n \leq p$. Then

$$\mathbb{E}\left[f(\mathbf{G})\right] \geq a_n - w(\Omega),$$

where a_n and $w(\cdot)$ are defined as in the escape-through-the-mesh theorem.

Combining this theorem and the Tsirelson-Ibragimov-Sudakov inequality, we obtain the escape-through-the-mesh theorem. Note that $\mathbf{G}$ is statistically equivalent to $\sqrt{n}\mathbf{A}$.

Probabilistic results for the *noiseless* case

Assume that $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a *matrix of i.i.d. Gaussian random variables* with zero means and variances $1/n$.

Let Ω be the intersection of $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ and the unit ℓ_2-norm sphere.

Theorem (Noiseless)

We have $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ *with probability at least* $1 - \exp\left\{-\frac{1}{2}\left[a_n - w(\Omega)\right]^2\right\}$, *provided that* $n \geq w(\Omega)^2 + 1$.

Proof.

Replace Ω by the intersection of $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ and the unit ℓ_2-norm sphere in the escape-through-the-mesh theorem. Note that the escape-through-the-mesh theorem is only meaningful when $a_n \geq w(\Omega)$; this condition leads to the constraint $n \geq w(\Omega)^2 + 1$. □

Probabilistic results for the *noisy* case

Assume that $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a *matrix of i.i.d. Gaussian random variables* with zero means and variances $1/n$.

Let Ω be the intersection of $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ and the unit ℓ_2-norm sphere.

Theorem (Noisy)

For any $\mu \in (0,1)$, *we have* $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\delta}{\sqrt{\mu}}$ *with probability at least* $1 - \exp\left\{-\frac{1}{2}\left[a_n - w(\Omega) - \sqrt{\mu n}\right]^2\right\}$ *provided that* $\|\mathbf{w}\|_2 \leq \delta$ *and* $n \geq \frac{w(\Omega)^2 + \frac{3}{2}}{(1-\sqrt{\mu})^2}$.

Proof.

Replace Ω by the intersection of $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ and the unit ℓ_2-norm sphere in the escape-through-the-mesh theorem. Note that the escape-through-the-mesh theorem is only meaningful when $a_n \geq w(\Omega) + \sqrt{\mu n}$; this condition leads to the constraint $n \geq \frac{w(\Omega)^2 + \frac{3}{2}}{(1-\sqrt{\mu})^2}$, assuming $\mu \in (0,1)$. □

Interpretation of the results

Recall the result in the previous slide.

Theorem (Noisy)

For any $\mu \in (0,1)$, *we have* $\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\kappa}{\sqrt{\mu}}$ *with probability at least* $1 - \exp\left\{-\frac{1}{2}\left[a_n - w(\Omega) - \sqrt{\mu n}\right]^2\right\}$ *provided that* $\|w\|_2 \leq \kappa$ *and* $n \geq \frac{w(\Omega)^2 + \frac{3}{2}}{(1-\sqrt{\mu})^2}$.

We have an equivalent formulation assuming $\kappa = \|\mathbf{w}\|_2$.

Theorem

For any $\mu \in (0,1)$, *we have*

$$\left\|\hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural\right\|_2 \leq \frac{2\sqrt{n}}{a_n - w(\Omega) - t}\|\mathbf{w}\|_2 \leq \frac{2\sqrt{n}}{\sqrt{n} - w(\Omega) - t}\|\mathbf{w}\|_2$$

with probability at least $1 - \exp\left(-\frac{1}{2}t^2\right)$ *provided* $n \geq \frac{w(\Omega)^2 + \frac{3}{2}}{(1-\sqrt{\mu})^2}$.

Observation: The quantity $w(\Omega)^2$ characterizes the degree of freedom of $\mathbf{x}^\natural$.

Remark: We will discuss an improvement of this guarantee.

Gaussian width

Definition (Gaussian width)

The Gaussian width $w(\Omega)$ of a set $\Omega \subset \mathbb{R}^n$ is given by

$$w(\Omega) := \max_{\mathbf{x}} \{\mathbb{E}\left[\langle \mathbf{g}, \mathbf{x} \rangle\right] : \mathbf{x} \in \Omega\},$$

where $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Example

Let V be a d-dimensional subspace of $\mathbb{R}^p$, and let Ω be the intersection of V and the unit ℓ_2-norm sphere. Then $w(\Omega) = \sqrt{d}$.

This justifies our claim that $[w(\Omega)]^2$ characterizes the degree of freedom of a set.

Proposition

1. *The Gaussian width is invariant under translation and unitary transforms (rotations).*
2. *Let $\mathcal{C}_1 \subseteq \mathcal{C}_2 \subseteq \mathbb{R}^n$. Then $w(\mathcal{C}_1) \leq w(\mathcal{C}_2)$.*

Examples

Let Ω always denote the intersection of $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ and the unit ℓ_2-norm sphere.

Example ([14])

1. Let $\mathcal{A} = \{\mathbf{e}_1, \dots, \mathbf{e}_p\}$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ with at most s non-zero entries. Then $\|\cdot\|_{\mathcal{A}}$ is the ℓ_1-norm, and $w(\Omega)^2 \leq 2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s$.
2. Let $\mathcal{A} = \{-1, +1\}^p$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ be a convex combination of k vectors in $\mathcal{A}$. Then $\|\cdot\|_{\mathcal{A}}$ is the ℓ_∞-norm, and $w(\Omega)^2 \leq \frac{p+k}{2}$.
3. Let $\mathcal{A} = \left\{\mathbf{X} : \operatorname{rank}(\mathbf{X}) = 1, \|\mathbf{X}\|_F = 1, \mathbf{X} \in \mathbb{R}^{p\times p}\right\}$, and let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with rank r. Then $\|\cdot\|_{\mathcal{A}}$ is the nuclear norm, and $w(\Omega)^2 \leq 3r(2p - r)$.

Some applications follow directly.

Application 1: Compressive sensing

Problem formulation [11, 19]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ with at most s non-zero entries, and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. How do we estimate $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes unknown noise?

Example

Let $\mathcal{A} = \{\mathbf{e}_1, \dots, \mathbf{e}_p\}$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ with at most s non-zero entries. Then $\|\cdot\|_{\mathcal{A}}$ is the ℓ_1-norm, and $w(\Omega)^2 \leq 2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s$.

Choose $\mathbf{A}$ to be a matrix of i.i.d. Gaussian random variables with zero means and variances $1/n$. Then by

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}$$

with $\kappa = \|\mathbf{w}\|_2$, we have

$$\left\| \hat{\mathbf{x}}_{\text{BPDN}} - \mathbf{x}^\natural \right\|_2 \lesssim \frac{2\sqrt{n}}{\sqrt{n} - \sqrt{2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s}} \|\mathbf{w}\|_2 .$$

Application 2: Multi-knapsack feasibility problem

Problem formulation [26]

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ which is a convex combination of k vectors in $\mathcal{A} := \{-1, +1\}^p$, and let $\mathbf{A} \in \mathbb{R}^{n \times p}$. How large should n be such that we can recover $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ via

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_\infty : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}?$$

Example

Let $\mathcal{A} = \{-1, +1\}^p$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ be a convex combination of k vectors in $\mathcal{A}$. Then $\|\cdot\|_{\mathcal{A}}$ is the ℓ_∞-norm, and $w(\Omega)^2 \leq \frac{p+k}{2}$.

Choose $\mathbf{A}$ to be a matrix of i.i.d. Gaussian random variables with zero means and variances $1/n$. Then we have

$$\mathbb{P}\left(\left\{ \hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural \right\} \right) \gtrsim 1 - \exp\left\{ -\frac{1}{2} \left[\sqrt{n} - \sqrt{\frac{p+k}{2}} \right]^2 \right\}.$$

Application 3: Matrix completion

Problem formulation [8, 18]

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with $\operatorname{rank}(\mathbf{X}^\natural) = r$, and let $\mathbf{A}_1, \dots, \mathbf{A}_n$ be matrices in $\mathbb{R}^{p\times p}$. How do we estimate $\mathbf{X}^\natural$ given $\mathbf{A}_1, \dots, \mathbf{A}_n$ and $b_i = \operatorname{Tr}\left(\mathbf{A}\mathbf{X}^\natural\right) + w_i$, $i = 1, \dots, n$, where $\mathbf{w} := (w_1, \dots, w_n)^T$ denotes unknown noise?

Example

Let $\mathcal{A} = \left\{\mathbf{X} : \operatorname{rank}(\mathbf{X}) = 1, \|\mathbf{X}\|_F = 1, \mathbf{X} \in \mathbb{R}^{p\times p}\right\}$, and let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ with rank r. Then $\|\cdot\|_{\mathcal{A}}$ is the nuclear norm, and $w(\Omega)^2 \leq 3r(2p-r)$.

Choose each A_i to be a matrix of i.i.d. Gaussian random variables with zero means and variances $1/n$. Then by

$$\hat{\mathbf{X}}_{\text{BPDN}} \in \arg\min_{\mathbf{X}\in\mathbb{R}^{p\times p}} \left\{ \|\mathbf{X}\|_* : \sum_{i=1}^{n} \left(b_i - \operatorname{Tr}\left(\mathbf{A}_i\mathbf{X}\right)\right)^2 \leq \kappa^2 \right\}$$

with some $\kappa = \|\mathbf{w}\|_2^2$, we have

$$\left\|\hat{\mathbf{X}}_{\text{BPDN}} - \mathbf{X}^\natural\right\|_2 \lesssim \frac{2\sqrt{n}}{\sqrt{n} - \sqrt{3r(2p-r)}} \|\mathbf{w}\|_2 .$$

Sharper bounds with oracle information

Suppose that we are able to set

$$\hat{\mathbf{x}}_{\text{BPDN,oracle}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\}.$$

Theorem ([29])

With probability at least $1-6\exp\left(-t^2/26\right)$, *we have*

$$\left\|\hat{\mathbf{x}}_{BPDN,oracle}-\mathbf{x}^\natural\right\|_2 \leq \left[\frac{w(\Omega)+t}{a_{n-1}}\right]\left[\frac{2\sqrt{n}}{a_n-w(\Omega)-t}\right]\|\mathbf{w}\|_2$$

for any $t>0$, *where* Ω *denotes the intersection of* $\mathcal{T}_{\|\cdot\|_{\mathcal{A}}}\left(\mathbf{x}^\natural\right)$ *and the unit* ℓ_2*-norm sphere.*

Observation: Recall that our analysis gives that with probability at least $1-\exp\left(-t^2/2\right)$,

$$\left\|\hat{\mathbf{x}}_{\text{BPDN,oracle}}-\mathbf{x}^\natural\right\|_2 \lesssim \left[\frac{2\sqrt{n}}{a_n-w(\Omega)-t}\right]\|w\|_2.$$

An improvement by the factor $\frac{w(\Omega)+t}{a_{n-1}} \leq 1$ appears assuming access of the oracle information $\|\mathbf{w}\|_2$.

Restricted isometry principle as an alternative approach

We have discussed the *restricted strong convexity condition*. In some problem settings the performance guarantee of $\hat{\mathbf{x}}_{\text{BPDN}}$ can be proved by the

restricted isometry property (RIP) condition.

Restricted isometry property

- In the preceding discourse we attempted to recover a fixed $\mathbf{x}^\natural$ so the recovery guarantees are referred to as for each.
- RIP gives a stronger guarantee, called for all. They hold for **any** $\mathbf{x}^\natural$.
- Can be illustrated by a game with two players: **Player 1** (**P1**) vs. **Player 2** (**P2**)

For all game

P1 will choose $\mathbf{A}$ and **P2** will choose any s-sparse $\mathbf{x}^\natural$, then **P1** will recover.

For each game

P2 will choose a s-sparse $\mathbf{x}^\natural$ and **P1** will choose $\mathbf{A}$, then **P1** will recover.

Recovery guarantee of BPDN with RIP

Given $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$, if $\mathbf{A}$ has RIP, then the ℓ_2-error between the BPDN solution, i.e. $\hat{\mathbf{x}}_{\text{BPDN}}$, and $\mathbf{x}^\natural$ is given by:

$$\|\mathbf{x}^\natural - \hat{\mathbf{x}}_{\text{BPDN}}\|_2 \leq C\|\mathbf{w}\|_2,$$

where $C > 0$ is a constant dependent only on RIP.

Restricted isometry property contd.

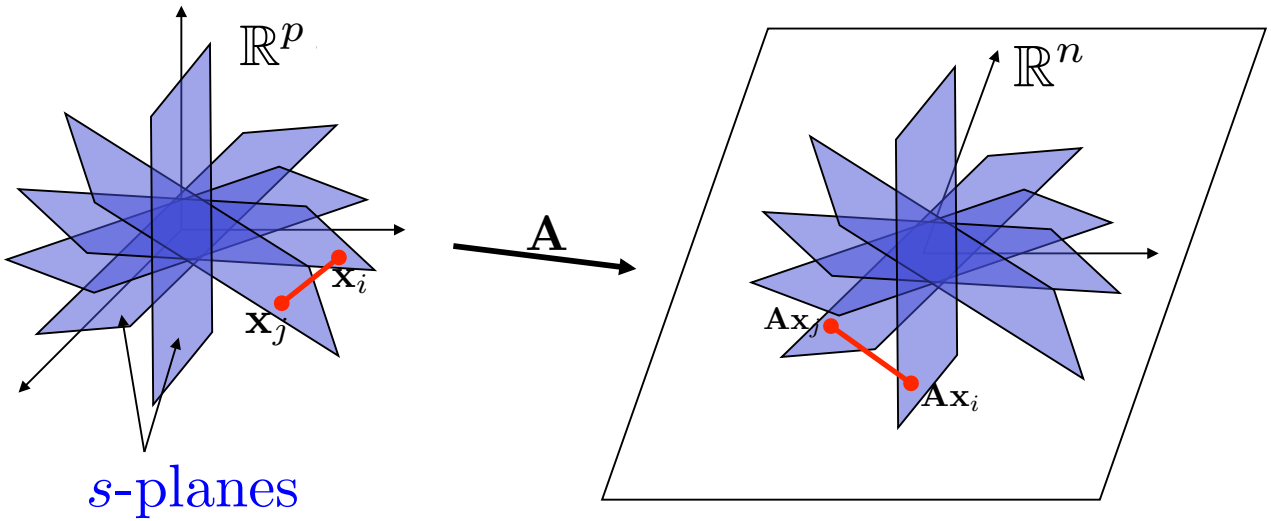
Definition (Restricted isometry constant [9])

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$. The s-th restricted isometry constant $\delta_s(\mathbf{A})$ of the matrix $\mathbf{A}$ is the smallest $\delta \geq 0$ such that

$$(1 - \delta) \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \delta) \|\mathbf{x}\|_2^2$$

for all s-sparse $\mathbf{x} \in \mathbb{R}^p$.

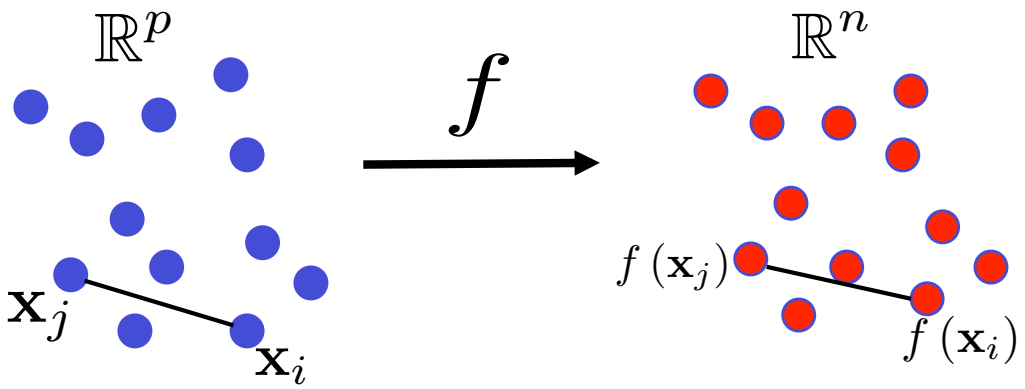
- RIP $\Rightarrow$ well-conditioning of submatrices of $\mathbf{A}$ restricted on s columns.



RIP and JL lemma

- RIP is related to Johnson-Lindenstraus (JL) lemma in *dimensionality reduction*.
- JL lemma is about pairwise distance preserving embedding of point clouds in a **high dimensional Euclidean space** into a much *lower dimensional space*.

Isometric embedding of m points in $\mathbb{R}^p$ into $\mathbb{R}^n$ for $n = \mathcal{O}(\log m)$



m points

Theorem (JL lemma)

Let $\varepsilon \in (0, 1/2)$ and let $\{\mathbf{x}_i\}_{i=1}^m \in \mathbb{R}^p$ be arbitrary points. For $n = \mathcal{O}\left(\varepsilon^{-2} \log m\right)$, there exists a map $f : \mathbb{R}^p \to \mathbb{R}^n$ such that

$$(1-\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \leq \|f(\mathbf{x}_i) - f(\mathbf{x}_j)\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \quad \forall i, j \in \{1, \ldots, m\}.$$

RIP and JL lemma contd.

Proof sketch for f being a random linear map.

- Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a random linear map and the set of pairwise differences $\mathcal{X} = \{\mathbf{x}_i - \mathbf{x}_j\} =: \{\mathbf{y}_l\}_{i=1}^{q}$ for $q = \binom{m}{2}$. Then we need to prove that

$$(1-\varepsilon)\,\|\mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{y}\|_2^2 \leq (1+\varepsilon)\,\|\mathbf{y}\|_2^2 \quad \forall \mathbf{y} \in \mathcal{X}.$$

- Since $\mathbf{A}$ is a random matrix, the proof that $\mathbf{A}$ satisfies the JL lemma with high probability (w.h.p) reduces to proving the concentration inequality:

$$\mathbb{P}\left[(1-\varepsilon)\,\|\mathbf{z}\|_2^2 \leq \|\mathbf{A}\mathbf{z}\|_2^2 \leq (1+\varepsilon)\,\|\mathbf{z}\|_2^2\right] \geq 1 - 2\exp\left(-c_0\varepsilon^2 n\right)$$

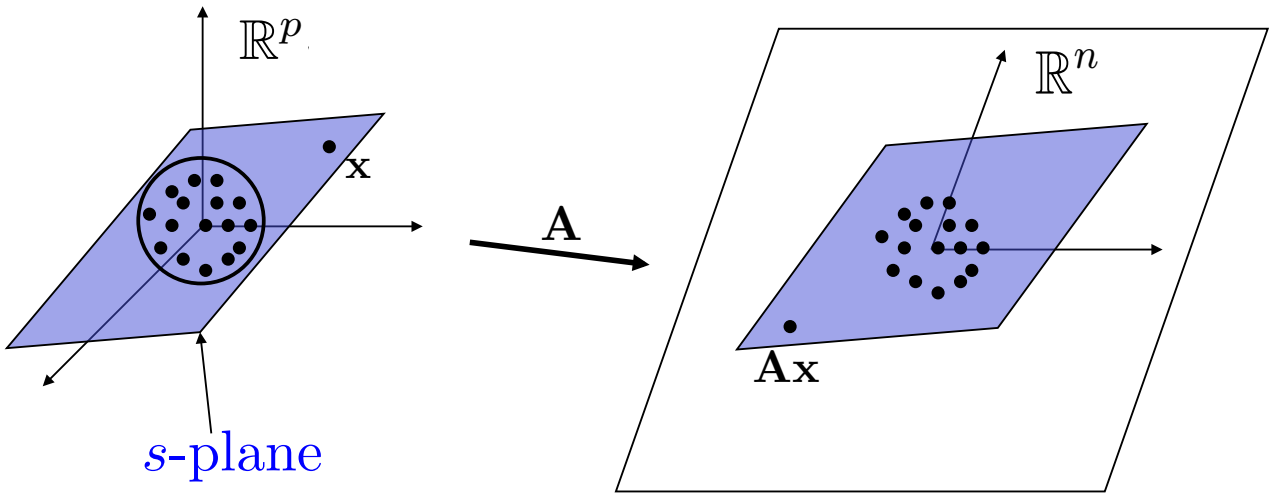
  for an arbitrary fixed $\mathbf{z} \in \mathbb{R}^p$, where c_0 is an absolute constant.

- The probability bound follows from a **union bound** over all $\binom{m}{2}$ $\mathbf{z} \in \mathcal{X}$.
- Choosing $n = \mathcal{O}\left(\varepsilon^{-2}\log m\right)$ is enough to make the bound $\geq 1/2$. □

RIP and JL lemma contd.

Consider the effect of a random linear JL map $\mathbf{A}$ on each s-plane:

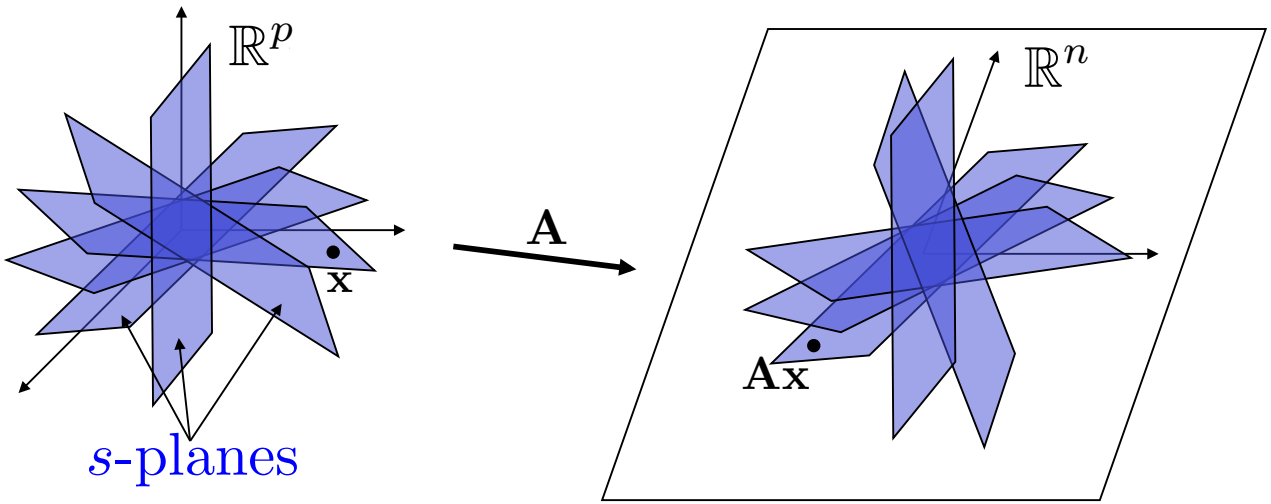
- construct covering of m points, $\mathcal{X}$, in a unit sphere.
- JL: isometry for each point with high probability.
- union bound $\Rightarrow$ isometry for all points $\mathbf{x}$ in $\mathcal{X}$.
- extends to isometry for all points $\mathbf{x}$ in s-plane



RIP and JL lemma contd.

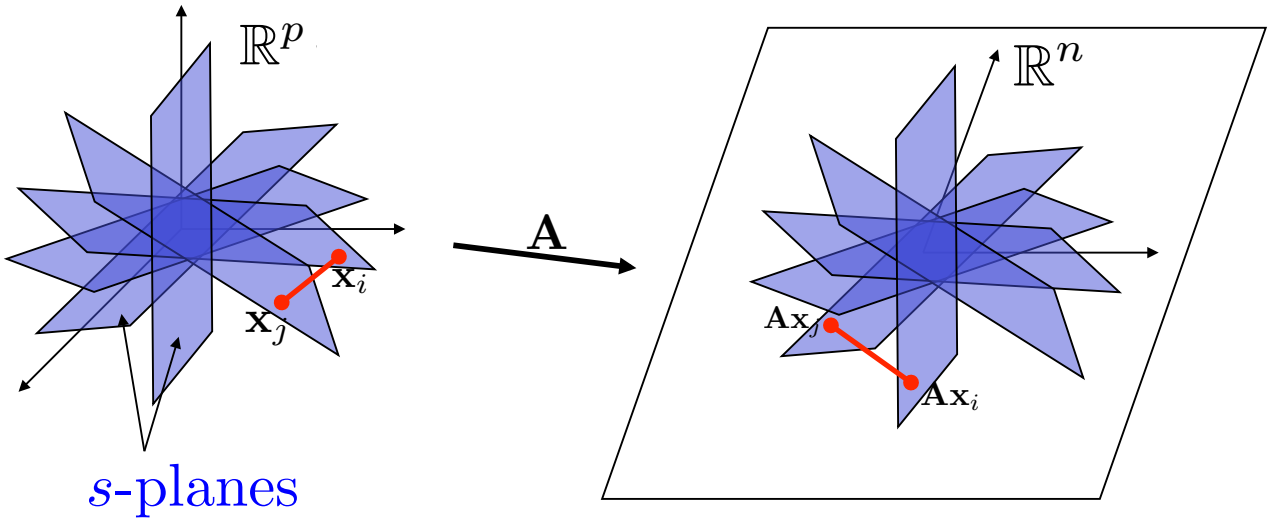
Consider the effect of a random linear JL map $\mathbf{A}$ on each s-plane:

- construct covering of m points, $\mathcal{X}$, in a unit sphere.
- JL: isometry for each point with high probability.
- union bound $\Rightarrow$ isometry for all points $\mathbf{x}$ in $\mathcal{X}$.
- extends to isometry for all points $\mathbf{x}$ in s-plane
- union bound $\Rightarrow$ isometry for all s-planes.



RIP and JL lemma contd.

RIP as a "stable" embedding



Theorem (RIP and sampling bounds)

For all s-sparse vectors $\mathbf{x} \in \mathbb{R}^p$, a random matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$ w.h.p satisfies RIP with a small δ_s if $n = \mathcal{O}\left(\delta_s^{-2} s \log(p/s)\right)$.

Corollary: Subgaussian matrices[2]

A subgaussian matrix (like Gaussian, Bernoulli, ...) in $\mathbb{R}^{n \times p}$ satisfies the RIP with high probability, if $n = \mathcal{O}(s \log(p/s))$.

[2] to be defined in Recitation 9

*Proof of RIP: The Algebra

- We show a more detailed proof of the RIP for a class of random matrices [2].
- The proof requires the lemma below which uses the distance preserving concentration inequality:

$$\mathbb{P}\left(\left|\|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2\right| \geq t\|\mathbf{x}\|_2^2\right) \leq 2e^{-c(t)n}, \tag{1}$$

for an arbitrary $\mathbf{x} \in \mathbb{R}^p$, where $c(t)$ is a function of a parameter $t \in (0,1)$.

Lemma ([2])

Let $\mathbf{A}$ be a random $n \times p$ matrix drawn according to any distribution that satisfies the concentration inequality (1). Then, for any set $\mathcal{S}$ with $|\mathcal{S}| = s \leq n$ and any $0 < \delta < 1$, we have

$$(1-\delta)\|\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{x}\|_2 \leq (1+\delta)\|\mathbf{x}\|_2, \quad \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}} \tag{2}$$

with probability at least

$$1 - 2(12/\delta)^s e^{-c(\delta/2)n}.$$

*Proof of RIP: The Algebra contd.

Proof of lemma

- We start by proving the upper bound in (2).
- Without loss of generality, we let $\|\mathbf{x}\|_2 = 1$ for all $\mathbf{x} \in \mathcal{X}_\mathcal{S}$.
- Choose a finite set $\mathcal{Y}_\mathcal{S}$ such that $\mathcal{Y}_\mathcal{S} \subseteq \mathcal{X}_\mathcal{S}$, $\|\mathbf{y}\|_2 = 1$ for all $\mathbf{y} \in \mathcal{Y}_\mathcal{S}$, and

$$\min_{\mathbf{y} \in \mathcal{Y}_\mathcal{S}} \|\mathbf{x} - \mathbf{y}\|_2 \leq \delta/4.$$

- From covering numbers such a set $\mathcal{Y}_\mathcal{S}$ exist with $|\mathcal{Y}_\mathcal{S}| \leq (12/\delta)^s$.
- Apply

$$\mathbb{P}\left(\left| \|\mathbf{A}\mathbf{x}\|_2^2 - \|\mathbf{x}\|_2^2 \right| \geq t \|\mathbf{x}\|_2^2 \right) \leq 2e^{-c(t)n},$$

  to $\mathcal{Y}_\mathcal{S}$ with $t = \delta/2$ by using a union bound.
- Then with probability $\geq 1 - 2(12/\delta)^s e^{-c(\delta/2)n}$ we have

$$(1 - \delta/2) \|\mathbf{y}\|_2^2 \leq \|\mathbf{A}\mathbf{y}\|_2^2 \leq (1 + \delta/2) \|\mathbf{y}\|_2^2, \quad \forall \mathbf{y} \in \mathcal{Y}_\mathcal{S}.$$

- It is easy to show that the above inequality reduces to

$$(1 - \delta/2) \|\mathbf{y}\|_2 \leq \|\mathbf{A}\mathbf{y}\|_2 \leq (1 + \delta/2) \|\mathbf{y}\|_2, \quad \forall \mathbf{y} \in \mathcal{Y}_\mathcal{S}.$$

*Proof of RIP: The Algebra contd.

Proof of lemma contd.

- Let α be the smallest number such that

$$\|\mathbf{A}\mathbf{x}\|_2 \leq (1+\alpha)\,\|\mathbf{x}\|_2 \quad \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}}$$

- The goal is to show that $\alpha \leq \delta$.
- Since $\|\mathbf{x}\|_2 = 1\ \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}}$ and $\|\mathbf{y}\|_2 = 1\ \forall \mathbf{y} \in \mathcal{Y}_{\mathcal{S}}\ \Rightarrow \|\mathbf{x}-\mathbf{y}\|_2 \leq \delta/4$, we have

$$\|\mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{A}\mathbf{y}\|_2 + \|\mathbf{A}(\mathbf{x}-\mathbf{y})\|_2 \leq 1 + \delta/2 + (1+\alpha)\delta/4.$$

- Thus by the definition α satisfies

$$\alpha \leq 1 + \delta/2 + (1+\alpha)\delta/4, \quad \Rightarrow \alpha \leq \frac{3}{4}\delta\,(1-\delta/4) \leq \delta.$$

- This concludes the proof for the upper bound in (2)
- The lower bound in (2) follows from the above since

$$\|\mathbf{A}\mathbf{x}\|_2 \geq \|\mathbf{A}\mathbf{y}\|_2 - \|\mathbf{A}(\mathbf{x}-\mathbf{y})\|_2 \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta.$$

□

*Proof of RIP: The Algebra contd.

Here is the final wrap up.

- For each s-dimensional space $\mathcal{X}_{\mathcal{S}}$, $\mathbf{A}$ fails to satisfy

$$(1-\delta)\left\|\mathbf{x}\right\|_2 \leq \left\|\mathbf{A}\mathbf{x}\right\|_2 \leq (1+\delta)\left\|\mathbf{x}\right\|_2, \quad \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}}$$

  with probability $\leq 2(12/\delta)^s e^{-c(\delta/2)n}$.

- Taking a union bound over all such subspaces, i.e., $\binom{p}{s} \leq (ep/s)^s$ upper bounds the failure probability by

$$2(ep/s)^s(12/\delta)^s e^{-c(\delta/2)n} = 2e^{-c(\delta/2)n+s[\log(ep/s)+\log(12/\delta)]}$$

- Thus, there exists $c_1 > 0$ such that whenever $n \geq c_1 \delta^{-2} s \log(p/s)$ the exponent on the right hand side of the above inequality $\leq -c_0 \delta^2 n$

□

RIP and other matrix ensembles

Partial Fourier

Description: Composed of subsampled rows of a Fourier matrix.

Application: Wireless communications, radar, etc

RIP: Requires $n = \mathcal{O}\left(s \log^3 s \log p\right)$ to satisfy RIP [12].

Advantage: These matrices possess a fast matrix application (i.e. FFT).

Disadvantage: These matrices are dense as such pose storage constraints.

Partial Circulant

Description: Composed of subsampled rows of a Circulant matrix.

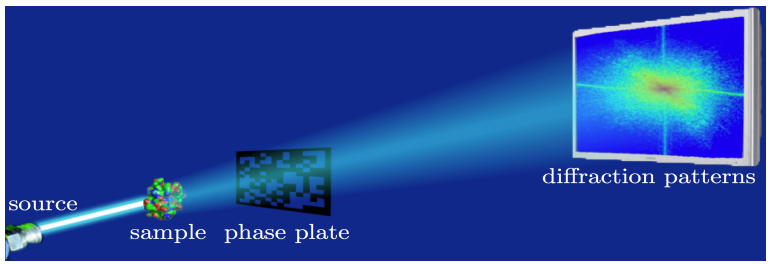
Application: Phase-retrieval

RIP: Requires $n = \mathcal{O}\left(\max\left\{s^{\frac{3}{2}} \log^{\frac{3}{2}} p, s \log^2 s \log^2 p\right\}\right)$ to satisfy RIP [30].

Advantage: These matrices are structured, hence easy to store and apply.

Disadvantage: These matrices are dense as such pose storage constraints.

Phase retrieval



Courtesy of [10]

RIP and other matrix ensembles contd.

Partial Toeplitz

Description: Composed of subsampled rows of a Toeplitz matrix.

Application: Sparse channel estimation

RIP: Requires $n = \mathcal{O}\left(s \log p + \log^3 p\right)$ to satisfy RIP [32].

Advantage: These matrices are structured, hence easy to store and apply.

Disadvantage: These matrices are dense as such pose storage constraints.

Deterministic matrices

Description: Constructed deterministically.

RIP: Requires $n = \mathcal{O}\left(s^{2-\nu}\right)$, for small $\nu > 0$ to satisfy RIP [7].

Advantage: These matrices can be implemented in hardware.

Disadvantage: The required number of measurements, n, is sub-optimal.

RIP and other matrix ensembles contd.

Binary $\{0,1\}^{n\times p}$ matrices

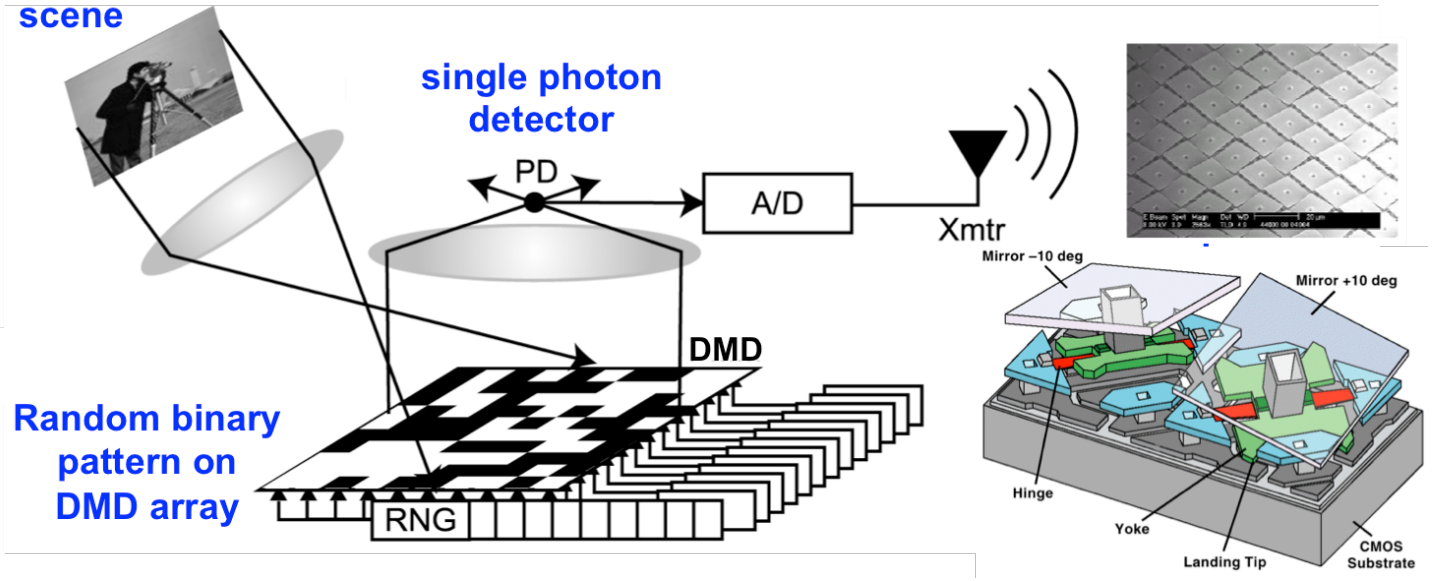
Description: Composed of subsampled rows of a Fourier matrix.

Application: Data streaming, graph sketching, single pixel camera, etc

RIP: Requires $n = \Omega\left(s^2\right)$ to satisfy RIP, $\Rightarrow$ square-root bottleneck [13].

Advantage: These matrices are sparse, hence easy to store and apply.

Disadvantage: The *square-root bottleneck*..



Recovery implications of the RIP

Recall the linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ and consider

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}.$$

Theorem (Noiseless [19])

Suppose that $\mathbf{x}^\natural$ is s-sparse and $\mathbf{w} = \mathbf{0}$, $\delta_{2s}(\mathbf{A}) < \frac{1}{3}$, and $\kappa = 0$, then $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$.

Theorem (Noisy [19])

Suppose that $\mathbf{x}^\natural$ is s-sparse, $\delta_{2s}(\mathbf{A}) < \frac{4}{\sqrt{41}}$, and $\|\mathbf{w}\|_2 \leq \kappa$, then

$$\left\| \hat{\mathbf{x}}_{BPDN} - \mathbf{x}^\natural \right\|_2 \leq c\kappa,$$

where c is some constant dependent only on $\delta_{2s}(\mathbf{A})$.

Generalization of the RIP

If $\mathbf{A}$ is sparse, computations involving $\mathbf{A}$ can be implemented more efficiently.

Bad news! A sparse matrix in $\mathbb{R}^{n\times p}$ *cannot* satisfy the RIP unless $n = \Omega(s^2)$ [13].

Good news! There exist a class of sparse $\mathbf{A}$ that satisfy a **variant of RIP** (defined below) with $n = \mathcal{O}\left(s\log(p/s)\right)$ [3].

Definition (RIP(q,s,δ) [3, 20])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$. The matrix $\mathbf{A}$ satisfies RIP(q,s,δ) of order s if for all s-sparse $\mathbf{x} \in \mathbb{R}^p$,

$$(1-\delta)\left\|\mathbf{x}\right\|_q \leq \left\|\mathbf{A}\mathbf{x}\right\|_q \leq \left\|\mathbf{x}\right\|_q.$$

Recall the linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural$ with s-sparse $\mathbf{x} \in \mathbb{R}^p$, and consider

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \left\|\mathbf{x}\right\|_1 : \left\|\mathbf{b}-\mathbf{A}\mathbf{x}\right\|_2 \leq \kappa \right\}.$$

Theorem ([3, 20])

There exists an $\mathbf{A} \in \mathbb{R}^{n\times p}$, for $\epsilon > 0$, with $n = O(s\log(p/s)/\epsilon^2)$ such that $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ and each column of $\mathbf{A}$ has $O(\log(n)/\epsilon)$ non-zero entries.

Main idea of the proof.

Prove the existence of such a matrix $\mathbf{A}$ that satisfies RIP$(1,s,\delta)$. □

RIP and low rank matrix recovery

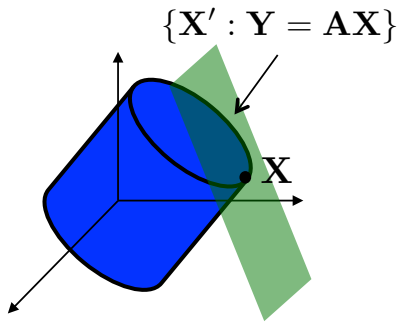
RIP is extended to give recovery guarantees for low rank matrix recovery [31].

Definition (RIP for matrices)

The linear map $\mathcal{A} : \mathbb{R}^{m\times p} \to \mathbb{R}^n$ satisfies RIP of order r if

$$(1-\delta_r(\mathcal{A}))\,\|\mathbf{X}\|_F^2 \le \|\mathcal{A}(\mathbf{X})\|_F^2 \le (1+\delta_r(\mathcal{A}))\,\|\mathbf{X}\|_F^2, \quad \forall \mathbf{X} : \mathrm{rank}(\mathbf{X}) \le r.$$

- RIP equivalent to bi-Lipschitz embedding of low-rank matrices.
- RIP guarantees a "stable" embedding of low-rank matrices.



Theorem (Measurement scaling)

Let $0 < \delta < 1$ *and* $\mathcal{A}(\mathbf{X}) = \mathbf{A}\mathrm{vex}(\mathbf{X})$ *where* $\mathbf{A} \in \mathbb{R}^{n\times mp}$ *and* $\mathrm{vec}(\mathbf{X})$ *vectorizes* $\mathbf{X}$. *If* $\mathbf{A}$ *is subgaussian and* $1 < r < n$, *then with probability* $1 - \exp(-c_0 n)$, $\delta_r(\mathcal{A}) \le \delta$ *whenever* $n \ge c_1 r \min(m,p)\log(mp)$ *for* $c_0, c_1 > 0$.

Coherence

Besides **RIP**, coherence is also used to give sparse recovery guarantees [19].

Definition (Coherence of a matrix)

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns $\{\mathbf{a}_i\}_{i=1}^p$. The *coherence* $\eta = \eta(\mathbf{A})$ of $\mathbf{A}$ is defined as

$$\eta := \max_{1 \le i \neq j \le p} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| .$$

Definition (ℓ_1-coherence)

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns. The *ℓ_1-coherence* η_1 of $\mathbf{A}$ is defined as

$$\eta_1(s) := \max_{1 \le i \le p} \max_{\mathcal{S}} \left\{ \sum_{j \in \mathcal{S}} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| , \ \mathcal{S} \subseteq [p], \ |\mathcal{S}| = s, \ i \notin \mathcal{S} \right\} .$$

- Generally, for $1 \le s, t \le p$ with $s + t \le p - 1$,

$$\max\{\eta_1(s), \eta_1(t)\} \le \eta_1(s+t) \le \eta_1(s) + \eta_1(t).$$

- Particularly, for $1 \le s \le p$, $\quad \eta \le \eta_1(s) \le s\eta$.

Coherence contd.

Theorem (Coherence and RIP)

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns, and let $1 \leq s \leq p$. For all s-sparse vectors $\mathbf{x} \in \mathbb{R}^p$,

$$(1 - \eta_1(s-1)) \|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1 + \eta_1(s-1)) \|\mathbf{x}\|_2^2.$$

- Equivalently, for each $\mathcal{S} \subseteq [p]$ with $|\mathcal{S}| \leq s$, the eigenvalues of $\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}}$ satisfy

$$1 - \eta_1(s-1) \leq \lambda \left(\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}} \right) \leq 1 + \eta_1(s-1).$$

- This means if $\eta_1(s-1) < 1$, then $\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}}$ is invertible.

Corollary

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns, and let $1 \leq s \leq p$. If

$$\eta_1(s) + \eta_1(s-1) < 1,$$

then for each $\mathcal{S} \subseteq [p]$ with $|\mathcal{S}| \leq 2s$, $\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}}$ is invertible and $\mathbf{A}_{\mathcal{S}}$ is injective.

This similarly holds if

$$\eta < (2s-1)^{-1}.$$

Coherence contd.

Theorem (Coherence bound)

The coherence and the ℓ_1-coherence of $\mathbf{A} \in \mathbb{R}^{n \times p}$ with ℓ_2-normalized columns respectively satisfies,

$$\eta \geq \sqrt{\frac{p-n}{n(p-1)}}, \quad \textit{and} \quad \eta_1(s) \geq s\sqrt{\frac{p-n}{n(p-1)}} \quad \textit{for} \quad s < \sqrt{p-1}.$$

Proposition (Coherence and RIP)

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a matrix with ℓ_2-normalized columns, then

$$\delta_1 = 0, \qquad \delta_2 = \eta, \qquad \delta_s \leq \eta_1(s-1) \leq \eta(s-1), \ s \geq 2.$$

Proof sketch.

To show that $\delta_s \leq \eta_1(s-1)$ we do the following:

1. Assuming that the columns of $\mathbf{A}$ are ℓ_2-normalized, we estimate $\lambda\left(\mathbf{A}_{\mathcal{S}}^T \mathbf{A}_{\mathcal{S}} - \mathbf{I}\right)$.
2. We then take the supremum over all $\mathcal{S} \subset [p]$ with $|\mathcal{S}| = s$, leading to the ℓ_1-coherence function $\eta_1(s-1)$.

□

Coherence contd.

- **Note** therefore that the estimation of the RIP constants of $\mathbf{A}$ relies on the estimation of $\lambda\left(\mathbf{A}_{\mathcal{S}}^T\mathbf{A}_{\mathcal{S}}-\mathbf{I}\right)$.
- But estimating these eigenvalues for a deterministic $\mathbf{A}$ relies on the Gershgorin's circle theorem:

Theorem (Gershgorin's circle theorem)

Let $\mathbf{A}\in\mathbb{R}^{p\times p}$ be a square matrix and let λ be an eigenvalue. Then there exists an index $j\in[p]$ such that

$$|\lambda-a_{jj}|\leq\sum_{i\in[p]\setminus\{j\}}|a_{ji}|\,.$$

- Recalling the δ^{-2} in the **sampling bound**, this implies that the coherence bound and the proposition above will result in the square root bollteneck in the sampling complexity of deterministic $\mathbf{A}$.

Coherence contd.

Recall the linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ and consider

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}.$$

Theorem (Noiseless)

Suppose that $\mathbf{x}^\natural$ *is* s*-sparse,* $\mathbf{w} = \mathbf{0}$*, and* $\kappa = 0$*, then* $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ *if*

$$\eta_1(s) + \eta_1(s-1) < 1, \quad \textit{or} \quad \eta < (2s-1)^{-1}.$$

Different formulations

Recall the basis pursuit denoising estimator

$$\hat{\mathbf{x}}_{\text{BPDN}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2 \leq \kappa \right\}.$$

There are two formulations closely related to $\hat{\mathbf{x}}_{\text{BPDN}}$.

Definition (Least absolute shrinkage and selection operator (lasso) [37])

$$\hat{\mathbf{x}}_{\text{lasso}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 : \|\mathbf{x}\|_1 \leq \tau \right\}.$$

Definition (Penalized least-squares)

$$\hat{\mathbf{x}}_{\text{PLS}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1 \right\}.$$

- $\hat{\mathbf{x}}_{\text{PLS}}$ is usually also called the lasso in literature.

We characterize the relations among the three formulations via the *Pareto curve*.

Pareto curve

Define $\hat{\mathbf{x}}_{\text{lasso}}(\tau) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 : \|\mathbf{x}\|_1 \leq \tau \right\}$.

Definition (Pareto curve)

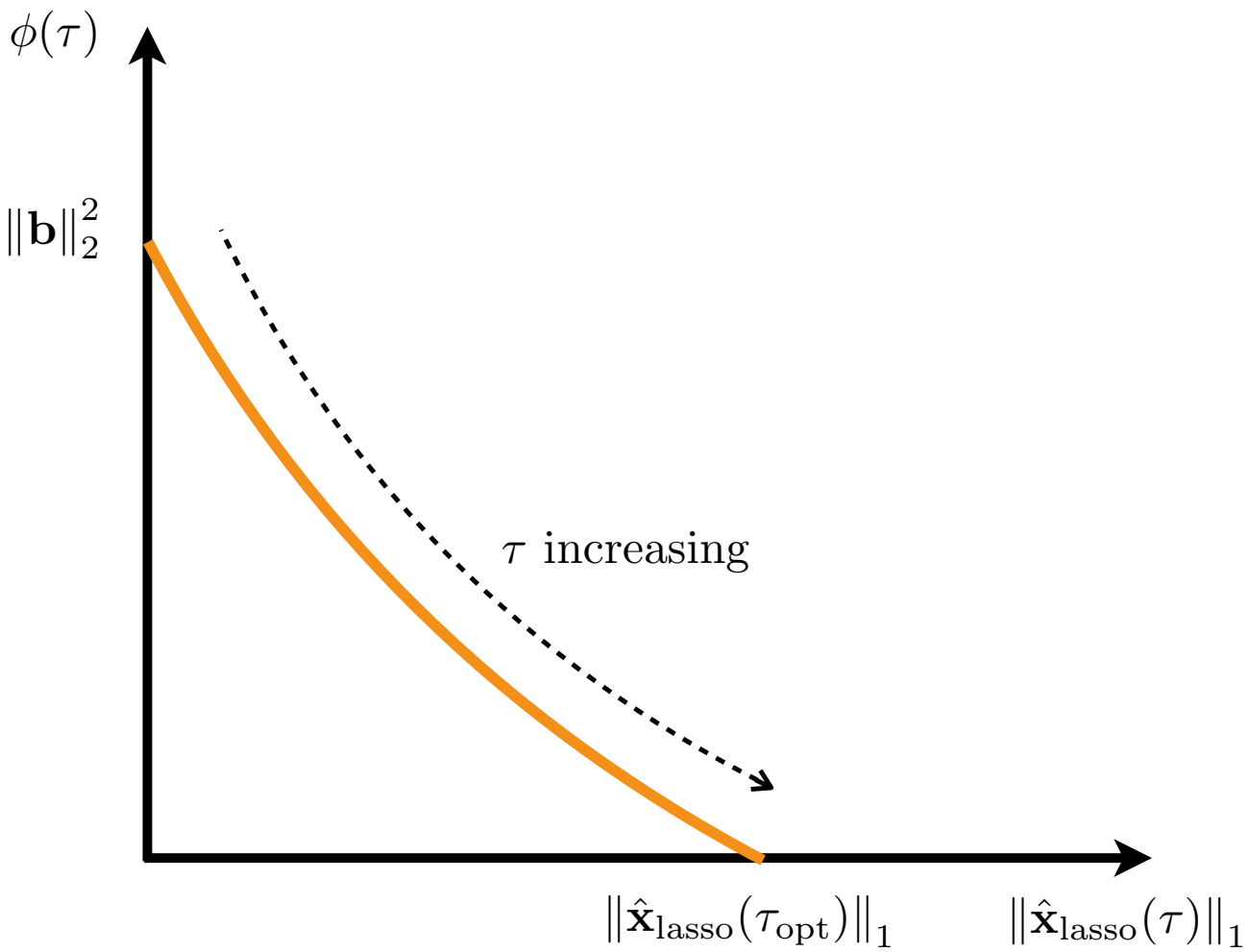
$$\phi(\tau) := \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{lasso}}(\tau)\|_2^2, \quad \tau \in [0, +\infty).$$

Theorem ([39])

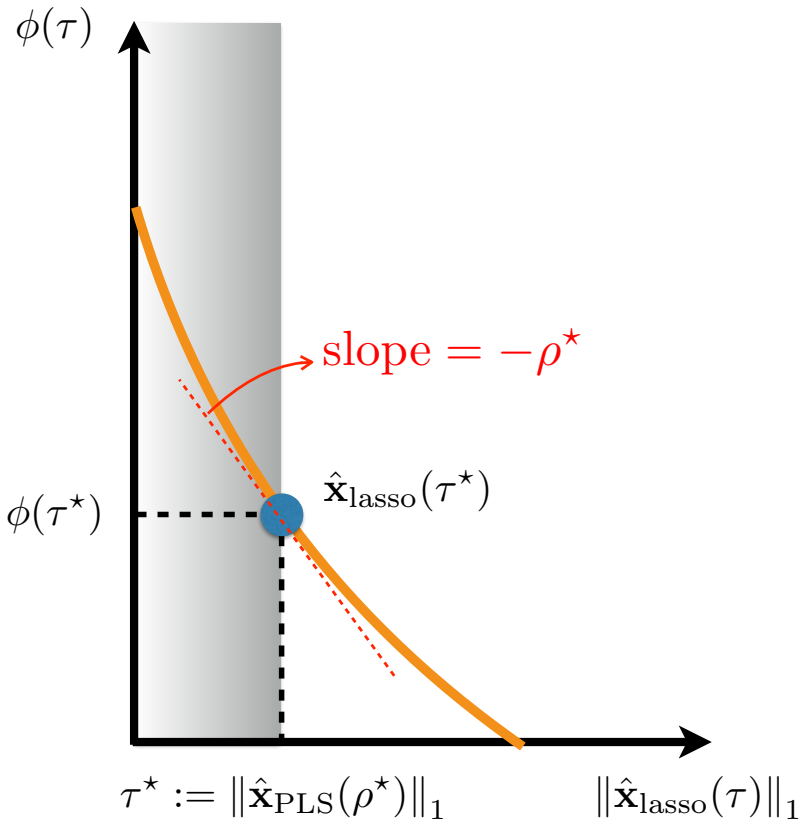
Define $\tau_{opt} := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1, \mathbf{b} = \mathbf{A}\mathbf{x} \right\}$.

1. *The function* $\phi(\tau)$ *is convex and nonincreasing.*
2. *The function* $\phi(\tau)$ *is strictly decreasing on* $[0, \tau_{opt}]$.
3. *The function* $\phi(\tau)$ *is continuously differentiable on* $(0, \tau_{opt})$.
4. *For all* $\tau \in [0, \tau_{opt}]$, $\|\hat{\mathbf{x}}_{lasso}(\tau)\|_1 = \tau$.

A typical Pareto curve



Relation between $\hat{\mathbf{x}}_{\text{PLS}}$ and $\hat{\mathbf{x}}_{\text{lasso}}$



Proposition

Let $\rho^\star > 0$ and $\tau^\star := \|\hat{\mathbf{x}}_{PLS}(\rho^\star)\|_1$. Then $\hat{\mathbf{x}}_{lasso}(\tau^\star) = \hat{\mathbf{x}}_{PLS}(\rho^\star)$.

Proof.

By definition, for all $\mathbf{x}$,

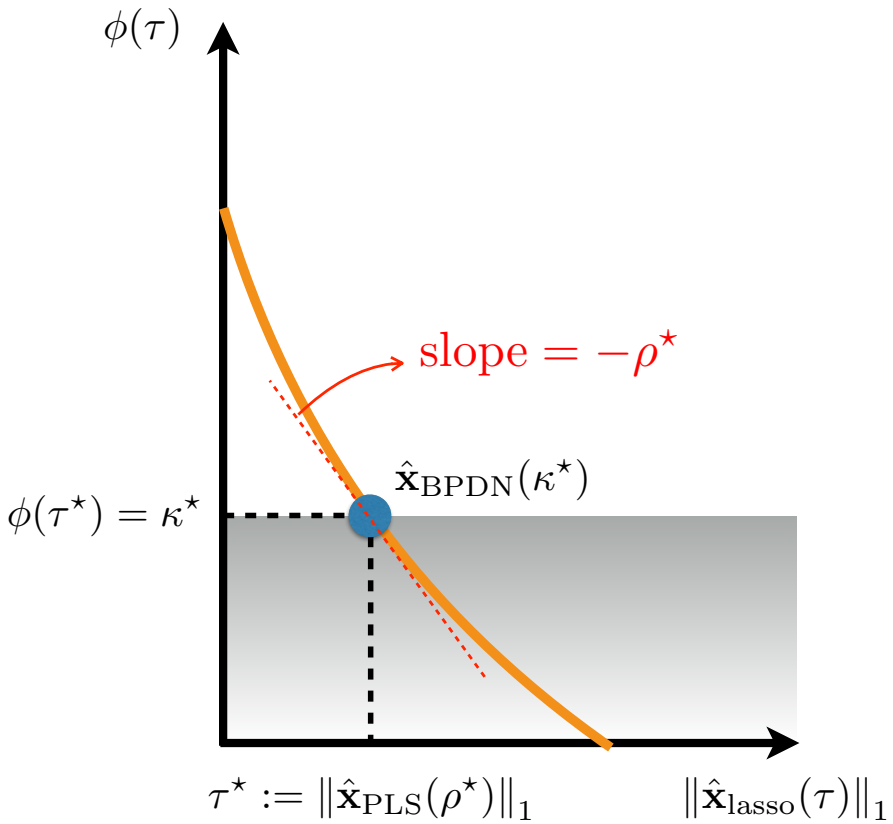
$$\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_2^2 + \rho^\star \|\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_1 \leq \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho^\star \|\mathbf{x}\|_1 .$$

Thus, for all $\mathbf{x}$ such that $\|\mathbf{x}\|_1 \leq \tau^\star$,

$$\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_2^2 \leq \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 .$$

□

Relation between $\hat{\mathbf{x}}_{\text{PLS}}$ and $\hat{\mathbf{x}}_{\text{BPDN}}$



Proposition

Let $\rho^\star > 0$ *and* $\kappa^\star := \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{PLS}(\rho^\star)\|_2$.
Then $\hat{\mathbf{x}}_{BPDN}(\kappa^\star) = \hat{\mathbf{x}}_{PLS}(\rho^\star)$.

Proof.

By definition, for all $\mathbf{x}$,

$$\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_2^2 + \rho^\star \|\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_1 \leq \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho^\star \|\mathbf{x}\|_1 .$$

Thus, for all $\mathbf{x}$ such that $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa^\star$,

$$\|\hat{\mathbf{x}}_{\text{PLS}}(\rho^\star)\|_1 \leq \|\mathbf{x}\|_1 .$$

□

Selection of the regularization parameter

Consider the Gaussian linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. We assume that there exists a matrix Ψ such that $\Psi\mathbf{x}^\natural$ is simple with respect to an atomic set $\mathcal{A}$, and we consider

$$\hat{\mathbf{x}}(\rho) = \arg\min_{\mathbf{x}} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_{\mathcal{A}} \right\}.$$

Problem

How do we choose the regularization coefficient ρ?

General principle

Fix a loss function $\mathcal{L}(\hat{\mathbf{x}}(\lambda), \mathbf{x}^\natural)$. Choose the ρ such that the risk $R(\rho) := \mathbb{E}\left[\mathcal{L}(\hat{\mathbf{x}}(\rho), \mathbf{x}^\natural)\right]$, or the expected loss, is minimized.

Issue

The risk $R(\rho)$ is intractable due to its dependence on $\mathbf{x}^\natural$. Thus it is impossible to find λ that minimizes the risk.

Some approaches

Popular approaches:

1. Covariance penalty
2. Cross validation
3. Upper bound heuristic

Common basic idea

Find a tractable estimate $\hat{R}(\rho)$ of the true risk $R(\rho)$. Choose $\rho^\star \in \arg\min_{\rho \geq 0} \hat{R}(\rho)$.

Covariance penalty

Recall the Gaussian linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w} \sim \mathcal{N}(\mathbf{A}\mathbf{x}^\natural, \sigma^2\mathbf{I})$ and we consider the case where $\|\cdot\|_{\mathcal{A}} := \|\cdot\|_1$,

$$\hat{\mathbf{x}}_{\text{PLS}}(\lambda) \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\Psi\mathbf{x}\|_1 \right\}.$$

Define the *expected prediction error* $R(\rho) := \mathbb{E}_{\mathbf{b},\tilde{\mathbf{b}}} \left[\left\| \tilde{\mathbf{b}} - \mathbf{A}\hat{\mathbf{x}}_{\text{PLS}}(\rho) \right\|_2^2 \right]$ with $\tilde{\mathbf{b}} \sim \mathcal{N}(\mathbf{A}\mathbf{x}^\natural, \sigma^2\mathbf{I})$ independent of $\mathbf{b}$.

Proposition ([25, 35, 16])

$$R(\rho) = \mathbb{E}_{\mathbf{b}} \left[\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{PLS}(\rho)\|_2^2 + 2\sigma^2 \mathrm{df} \right].$$

where $\mathrm{df} := \sigma^{-2}\mathrm{Tr}\left(\mathrm{cov}\left(\mathbf{A}\hat{\mathbf{x}}_{PLS}(\rho), \mathbf{b}\right)\right)$ *is called the degrees-of-freedom.*

Proof.

Note that $R(\rho) = \mathbb{E}_{\mathbf{b},\tilde{\mathbf{b}}} \left[\left\| \tilde{\mathbf{b}} - \mathbf{b} + \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\text{PLS}}(\rho) \right\|^2 \right]$. □

[3] The covariance penalty approach is also called Mallows' C_p approach (cf., a primitive version in [25]).

Covariance penalty

Basic idea of the covariance penalty approach

Let $\hat{R}(\rho)$ be an estimator of $R(\rho)$ such that $\mathbb{E}\left[\hat{R}(\rho)\right] = R(\rho)$. Choose $\rho^\star := \arg\min_\lambda \hat{R}(\rho)$.

Definition (Stein's unbiased risk estimator [35])

Any estimator $\hat{R}(\rho)$ of $R(\rho)$ such that $\mathbb{E}_{\tilde{\mathbf{b}}}\left[\hat{R}(\rho)\right] = R(\rho)$ is called a Stein's unbiased risk estimator (SURE).

Recall

$$R(\rho) = \mathbb{E}_{\mathbf{b}}\left[\|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho)\|_2^2 + 2\sigma^2 \mathrm{df}(\rho)\right],$$

Let $\widehat{\mathrm{df}}$ be an estimator of df such that $\mathbb{E}\left[\widehat{\mathrm{df}}\right] = \mathrm{df}$. Then

$$\hat{R}(\rho) := \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho)\|_2^2 + 2\sigma^2 \widehat{\mathrm{df}}$$

is a SURE of $R(\rho)$.

Covariance penalty

Theorem ([38, 40])

Define $\mathcal{S}(\rho)$ as the support set of $\hat{\mathbf{x}}_{PLS}(\rho)$, and let $\Psi_{\mathcal{S}(\rho)^c}$ consist of columns of Ψ that are not indexed by elements in $\mathcal{S}(\rho)$.

$$\mathrm{df} = \mathbb{E}\left[\dim\left(\left\{\mathbf{Az} : \mathbf{z} \in \mathrm{null}(\Psi_{\mathcal{S}(\rho)^c})\right\}\right)\right].$$

Thus we may choose $\widehat{\mathrm{df}} := \dim\left(\left\{\mathbf{Az} : \mathbf{z} \in \mathrm{null}(\Psi_{\mathcal{S}(\rho)^c})\right\}\right)$, and

$$\hat{R}(\rho) := \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\mathrm{PLS}}(\rho)\|_2^2 + \dim\left(\left\{\mathbf{Az} : \mathbf{z} \in \mathrm{null}(\Psi_{\mathcal{S}(\rho)^c})\right\}\right)$$

is a SURE of the risk $R(\rho)$.

Cross validation

A typical instance of the cross validation approach is the following one.

Leave-one-out cross validation [1, 5]

Let $\hat{\mathbf{x}}^{(-k)}(\rho)$ be an estimator based on $\mathbf{b}^{(-k)} := (b_1, \dots, b_{k-1}, b_{k+1}, \dots, b_n)^T$ and parameterized by $\rho > 0$, and let $\mathbf{A}^{(-k)} \in \mathbb{R}^{(n-1)\times p}$ be the matrix obtained by removing the kth row of $\mathbf{A}$. Define

$$\hat{R}(\rho) := \frac{1}{n} \sum_{k=1}^{n} \left[b_k - \mathbf{A}^{(-k)} \hat{\mathbf{x}}^{(-k)}(\rho) \right]^2 .$$

Choose $\rho^\star := \arg\min_{\rho>0} \hat{R}(\rho)$.

Remarks

A list of variants of the leave-one-out scheme can be found in [1].

Generalized cross validation

Definition (Ridge regression estimator)

$$\hat{\mathbf{x}}_{\text{ridge}}(\rho) \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_2^2 \right\}.$$

- There always exists a $\rho > 0$ such that $\mathbb{E}\left[\left\|\hat{\mathbf{x}}_{\text{ridge}}(\rho) - \mathbf{x}^\natural\right\|_2^2\right] \leq \mathbb{E}\left[\left\|\hat{\mathbf{x}}_{\text{LS}} - \mathbf{x}^\natural\right\|_2^2\right]$.

Generalized cross validation [21]

For ridge regression, the generalized cross validation approximates the leave-one-out approach.

$$\rho^\star := \arg\min_{\lambda\geq 0} \left\{ \frac{\frac{1}{n}\|\mathbf{b}-\mathbf{M}(\rho)\mathbf{b}\|}{\left[\frac{1}{n}\text{Tr}\left(\mathbf{I}-\mathbf{M}(\rho)\right)\right]^2} \right\},$$

where

$$\mathbf{M}(\rho) := \mathbf{A}(\mathbf{A}^T\mathbf{A} + n\rho\mathbf{I})^{-1}\mathbf{A}^T.$$

- There does not exist any explicit result about applying the cross validation method to the lasso currently.

Solution path

Consider $\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho) := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_1 \right\}$.

Definition (Solution path)

The *solution path* of $\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho)$ is the set $\{\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho) : \rho > 0\}$.

Recall that in the covariance penalty approach we aim at minimizing

$$\hat{R}_{\mathsf{CP}} := \|\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_{\mathsf{PLS}}(\rho)\|_2^2 + 2\sigma^2 \widehat{\mathrm{df}},$$

and in the cross validation approach we aim at minimizing

$$\hat{R}_{\mathsf{CV}}(\rho) := \frac{1}{n} \sum_{k=1}^{n} \left[b_k - \mathbf{A}^{(-k)} \hat{\mathbf{x}}_{\mathsf{PLS}}^{(-k)}(\rho) \right]^2 .$$

In both approaches we have to solve for the *solution path*.

Homotopy method

Theorem ([17, 24])

Let $\mathcal{S}(\rho)$ be the support of $\hat{\mathbf{x}}_{PLS}(\rho)$ for $\rho > 0$. Assume that for any $\rho > 0$, the submatrix $\left(\mathbf{A}^T\mathbf{A}\right)_{\mathcal{S}(\rho),\mathcal{S}(\rho)}$ is positive definite. Then the solution path is well defined, unique, continuous, and piecewise linear.

Insight: It suffices to find the *kinks*, or the points where the direction of the solution path changes, to characterize the whole solution path.

A *homotopy method* finds the pairs $(\rho_k, \hat{\mathbf{x}}_{\text{PLS}}(\rho_k))$ where ρ_k are the kinks (cf. [17, 28] for details).

Theorem ([24])

In the worst case the solution path can have exactly $(3^p + 1)/2$ kinks.

Insight: The computational complexity increases exponentially with p in the worst case, since to determine a kink we have to solve at least one lasso problem.

Good news: In practice we seldom encounter the worst case; the number of kinks is usually $O(p)$ by experience [33].[4]

[4] We will observe a similar gap between practical and worst-case performances in Lecture 9 for the simplex method.

Upper bound heuristic

Consider the linear model $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$ and we assume that $\mathbf{x}^\natural \in \mathbb{R}^p$ satisfies $\left\|\mathbf{x}^\natural\right\|_0 = s$ with some $s \leq p$ and $\mathbf{A} \in \mathbb{R}^{n\times p}$ is a matrix of i.i.d. random variables $\sim \mathcal{N}(0, 1/n)$.

$$\hat{\mathbf{x}}_{\mathsf{PLS}}(\lambda) \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \rho \|\mathbf{x}\|_1 \right\}.$$

Theorem ([36])

Assume that $n \geq 2$. Then for any $t \in (0, \sqrt{p-1} - \sqrt{c_{\mathbf{x}^\natural}}]$, with probability at least $1 - 5\exp(-t^2/32)$,

$$\left\|\hat{\mathbf{x}}_{PLS} - \mathbf{x}^\natural\right\|_2 \leq 2\|\mathbf{w}\| \frac{\sqrt{c_{\mathbf{x}^\natural}} + t}{\sqrt{n-1} - \sqrt{c_{\mathbf{x}^\natural}} - t},$$

where

$$c_{\mathbf{x}^\natural} := s(1+n\rho^2) + (p-s)\left[(1+n\rho^2)\mathrm{erfc}\left(\rho\sqrt{\frac{n}{2}}\right) - \sqrt{\frac{2n}{\pi}}\rho\exp\left(-\frac{n\rho^2}{2}\right)\right],$$

$\mathrm{erfc}(\cdot)$ *being the standard complementary error function.*

Upper bound heuristic

Observation: $c_{\mathbf{x}^\natural}$ is a function of n, p, s, and ρ only, and thus we have $\left\|\hat{\mathbf{x}}_{\text{PLS}} - \mathbf{x}^\natural\right\|_2 \leq \|\mathbf{w}\|_2 f(n, p, s, \rho, t)$. Note that f *only depends on* $s := \left\|\mathbf{x}^\natural\right\|_0$ instead of $\mathbf{x}^\natural$.

Lower bound heuristic [36]

Consider $\|\mathbf{w}\|_2 f(n, p, s, \rho, t)$ as an estimate of $\left\|\hat{\mathbf{x}}_{\text{PLS}} - \mathbf{x}^\natural\right\|_2$.

Suppose that $s = \left\|\mathbf{x}^\natural\right\|_0$ *is known.* Choose a $\rho^\star$ that minimizes $f(n, p, s, \rho, t)$ for a given set of n, p, s, t.

References

[1] Sylvain Arlot and Alain Celisse.
A survey of cross-validation procedures for model selection.
Stat. Surv., 4:40–79, 2010.

[2] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin.
A simple proof of the restricted isometry property for random matrices.
Constructive Approximation, 28(3):253–263, 2008.

[3] R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss.
Combining geometry and combinatorics: a unified approach to sparse signal recovery.
In *46th Annu. Allerton Conf.*, pages 798–805, September 2008.

[4] Peter Bickel, Ya'acov Ritov, and Alexandre B. Tsybakov.
Simultaneous analysis of Lasso and Dantzig selector.
Ann. Stat., 37(4):1705–1732, 2009.

[5] Peter J. Bickel and Bo Li.
Regularization in statistics.
Test, 15(2):271–344, 2006.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart.
Concentration Inequalities: A Nonasymptotic Theory of Independence.
Oxford Univ. Press, Oxford, 2013.

References

[7] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin, Denka Kutzarova, et al.
Explicit constructions of rip matrices and related problems.
Duke Mathematical Journal, 159(1):145–185, 2011.

[8] Emmanuel Candès and Benjamin Recht.
Exact matrix completion via convex optimization.
Found. Comput. Math., 9:717–772, 2009.

[9] Emmanuel J. Candès.
The restricted isometry property and its implications for compressed sensing.
C. R. Acad. Sci. Paris, Ser. I, 346:589–592, 2008.

[10] Emmanuel J Candès, Xiaodong Li, and Mahdi Soltanolkotabi.
Phase retrieval from coded diffraction patterns.
preprint, 2013.

[11] Emmanuel J. Candès, Justin Romberg, and Terence Tao.
Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.
IEEE Trans. Inf. Theory, 52(2):489–509, February 2006.

[12] Emmanuel J Candes and Terence Tao.
Near-optimal signal recovery from random projections: Universal encoding strategies?
Information Theory, IEEE Transactions on, 52(12):5406–5425, 2006.

References

[13] Venkat Chandar.

A negative result concerning explicit matrices with the restricted isometry property.

Technical report, 2008.

[14] 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.

[15] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders.

Atomic decomposition by basis pursuit.

SIAM J. Sci. Comput., 20(1):33–61, 1998.

[16] Bradley Efron.

The estimation of prediction error: Covariance penalities and cross-validation.

J. Am. Stat. Assoc., 99(467):619–632, September 2004.

[17] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani.

Least angle regression.

Ann. Stat., 32(2):407–499, 2004.

[18] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert.

Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators.

New J. Phys., 14, 2012.

References

[19] Simon Foucart and Holger Rauhut.
A Mathematical Introduction to Compressive Sensing.
Birkhäuser, Basel, 2013.

[20] Anna Gilbert and Piotr Indyk.
Sparse recovery using sparse matrices.
Proc. IEEE, 98(6):937–947, June 2010.

[21] Gene H. Golub, Michael Heath, and Grace Wahba.
Generalized cross-validation as a method for choosing a good ridge parameter.
Technometrics, 21(2):215–223, May 1979.

[22] Y. Gordon.
On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$.
In Joram Lindenstrauss and Vitali D. Milman, editors, *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, Berlin, 1988. Springer-Verl.

[23] Rémi Gribonval, Volkan Cevher, and Mike E. Davies.
Compressible distributions for high-dimensional statistics.
IEEE Trans. Inf. Theory, 58(8):5016–5034, 2012.

[24] Julien Mairal and Bin Yu.
Complexity analysis of the Lasso regularization path.
In *Proc. 29th Int. Conf. Machine Learning*, 2012.

References

[25] C. L. Mallows.

Some comments on C_p.

Technometrics, 15(4):661–675, November 1973.

[26] O. L. Mangasarian and Benjamin Recht.

Probability of unique integer solution to a system of linear equations.

Eur. J. Oper. Res., 214:27–30, 2011.

[27] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu.

A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers.

Stat. Sci., 27(4):538–557, 2012.

[28] M. R. Osborne, Brett Presnell, and B. A. Turlach.

A new approach to variable selection in least squares problems.

IMA J. Numer. Anal., 20:389–404, 2000.

[29] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi.

Simple bounds for noisy linear inverse problems with exact side information.

2013.

arXiv:1312.0641v2 [cs.IT].

References

[30] Holger Rauhut, Justin Romberg, and Joel A Tropp.
Restricted isometries for partial random circulant matrices.
Applied and Computational Harmonic Analysis, 32(2):242–254, 2012.

[31] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo.
Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization.
SIAM review, 52(3):471–501, 2010.

[32] Justin Romberg.
Compressive sensing by random convolution.
SIAM Journal on Imaging Sciences, 2(4):1098–1128, 2009.

[33] Saharon Rosset and Ji Zhu.
Piecewise linear regularized solution paths.
Ann. Stat., 35(3):1012–1030, 2007.

[34] Mark Rudelson and Roman Vershynin.
On sparse reconstruction from Fourier and Gaussian measurements.
Commun. Pure Appl. Math., LXI:1025–1045, 2008.

[35] Charles M. Stein.
Estimation of the mean of a multivariate normal distribution.
Ann. Stat., 9(6):1135–1151, 1981.

References

[36] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi.
Simple error bounds for regularized noisy linear inverse problems.
2014.
arXiv:1401.6578v1 [math.OC].

[37] Robert Tibshirani.
Regression shrinkage and selection via the lasso.
J. R. Stat. Soc., Ser. B, 58(1):267–288, 1996.

[38] Ryan J. Tibshirani and Jonathan Taylor.
Degrees of freedom in lasso problems.
Ann. Stat., 40(2):1198–1232, 2012.

[39] Ewout van den Berg and Michael P. Friedlander.
Probing the Pareto frontier for basis pursuit solutions.
SIAM J. Sci. Comput., 31(2):890–912, 2008.

[40] Hui Zou, Trevor Hastie, and Robert Tibshirani.
On the "degrees of freedom" of the lasso.
Ann. Stat., 35(5):2173–2192, 2007.