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)

License Information for Mathematics of Data Slides

Outline

- Today
  1. Source separation problem
  2. Incoherence and uncertainty principle
  3. General recipe for source separation
  4. Phase transition via statistical dimension
  5. Phase transition via convex polytopes
  6. Selection of the parameter
  7. Nonsmooth convex minimization by smoothing
- Next week
  1. Constrained convex minimization

Recommended reading

- D. Amelunxen *et al.*, "Living on the edge: Phase transitions in convex programs with random data," 2014, arXiv:1303.6672v2 [cs.IT].
- M.B. McCoy *et al.*, "Convexity in source separation," *IEEE Sig. Process. Mag.*, vol. 31, pp. 87–95, 2014.
- *D.L. Donoho and J. Tanner, "Counting faces of randomly projected polytopes when the projection radically lowers dimension," *J. Amer. Math. Soc.*, vol. 22, no. 1, pp. 1–53, 2009.
- Y. Nesterov, "Smooth minimization of nonsmooth functions," *Math. Program., Ser. A*, vol. 103, pp. 127–152, 2005.

Motivation

Motivation

This lecture illustrates how compressive sensing generalizes as a *source separation problem* in a unified framework.

It turns out that the formulation of a convex estimator for the source separation problem, in general, requires minimizing the sum of two *nonsmooth* convex functions. We derive the statistical performance guarantee of such an estimator, and show algorithms that address the composite nonsmooth convex minimization problems.

Outline

- Today
  1. *Source separation problem*
  2. Incoherence and uncertainty principle
  3. General recipe for source separation
  4. Phase transition via statistical dimension
  5. Phase transition via convex polytopes
  6. Nonsmooth convex minimization by smoothing
- Next week
  1. Constrained convex minimization

Source separation

Problem (Source separation)

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ *be two unknown vectors. How do we estimate* $\mathbf{x}^\natural$ *and* $\mathbf{y}^\natural$ *given* $\mathbf{z} := \mathbf{x}^\natural + \mathbf{y}^\natural$*?*

Source separation

Problem (Source separation)

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ *be two unknown vectors. How do we estimate* $\mathbf{x}^\natural$ *and* $\mathbf{y}^\natural$ *given* $\mathbf{z} := \mathbf{x}^\natural + \mathbf{y}^\natural$*?*

Observation

Source separation is impossible if we do not have any *additional information* about $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$.

Example

Obviously, without any additional information, the equation $\mathbf{z} = \mathbf{x}^\natural + \mathbf{y}^\natural$ has infinite possible solutions for $(\mathbf{x}^\natural, \mathbf{y}^\natural)$.

Insights from nearly trivial examples

Example

Let $\mathbf{z} = (2, 1)^T := \mathbf{x}^\natural + \mathbf{y}^\natural$. Without additional information it is impossible to perfectly recover $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$.

However, suppose now we know $\mathbf{x}^\natural = (x^\natural, 0)^T$ and $\mathbf{y}^\natural = (0, y^\natural)^T$, then we can perfectly recover $\mathbf{x}^\natural = (2, 0)^T$ and $\mathbf{y}^\natural = (0, 1)^T$.

Insight: To have a well-posed source separation problem, some information on the *signal structures* is needed.

Example

Suppose now that we know $\mathbf{x}^\natural = (2, x^\natural)^T$ and $\mathbf{y}^\natural = (0, y^\natural)^T$, then it is still impossible to perfectly recover $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$.

Insight: The structures must be *incoherent* in some sense.

A classical well-posed source separation problem

Problem (Spikes and sines)

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ *be sparse, and let* $\mathbf{D}$ *denote the discrete cosine transform (DCT) matrix. How do we estimate* $\mathbf{x}^\natural$ *and* $\mathbf{y}^\natural$ *given* $\mathbf{z} := \mathbf{x}^\natural + \mathbf{D}\mathbf{y}^\natural$*?*

$\mathbf{z}$

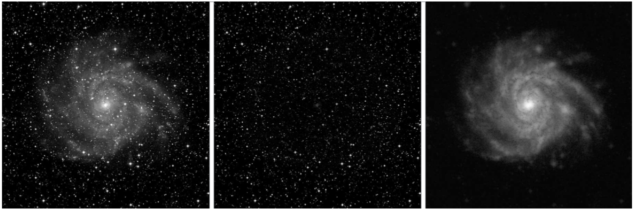
A classical well-posed source separation problem

Problem (Spikes and sines)

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ be sparse, and let $\mathbf{D}$ denote the discrete cosine transform (DCT) matrix. How do we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{D}\mathbf{y}^\natural$?

Observation: $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ are $\underbrace{\text{sparse}}_{\textit{signal structure}}$ $\underbrace{\text{in different bases}}_{\textit{incoherence}}$.

$$\mathbf{z} \quad = \quad \mathbf{x}^\natural \quad + \quad \mathbf{y}^\natural$$



Other applications of the source separation problem

Problem (Signal denoising [22])

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and let $\mathbf{w}^\natural \in \mathbb{R}^p$ denote some unknown noise. How do we estimate $\mathbf{x}^\natural$ (and thus also $\mathbf{w}^\natural$) given $\mathbf{b} = \mathbf{x}^\natural + \mathbf{w}^\natural$?

Applications: Wireless communications with narrowband interferences, signal processing with impulse noises, etc.

Problem (Morphological component analysis [11])

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ be sparse, and $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times p}$. How do we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{U}\mathbf{x}^\natural + \mathbf{V}\mathbf{y}^\natural$?

Applications: Spikes and Sines, texture separation, image inpainting, etc.

Problem (Robust principal component analysis (PCA) [3])

Let $\mathbf{X}^\natural \in \mathbb{R}^{p\times p}$ be sparse and $\mathbf{Y}^\natural \in \mathbb{R}^{p\times p}$ be low-rank. How do we estimate $\mathbf{X}^\natural$ and $\mathbf{Y}^\natural$ given $\mathbf{Z} := \mathbf{X}^\natural + \mathbf{Y}^\natural$?

Applications: Background separation in videos, face recognition, etc.

Outline

- Today
    1. Source separation problem
    2. *Incoherence and uncertainty principle*
    3. General recipe for source separation
    4. Phase transition via statistical dimension
    5. Phase transition via convex polytopes
    6. Nonsmooth convex minimization by smoothing
- Next week
    1. Constrained convex minimization

How do we solve the spikes and sines problem?

Problem (Spikes and sines)

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ be sparse, and let $\mathbf{D}$ denote the discrete cosine transform (DCT) matrix. How do we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{D}\mathbf{y}^\natural$?

We want to find sparse estimates $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ such that $\mathbf{z} = \hat{\mathbf{x}} + \mathbf{D}\hat{\mathbf{y}}$.

ℓ_0-"norm" approach

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 + \rho \|\mathbf{y}\|_0 : \mathbf{z} = \mathbf{x} + \mathbf{D}\mathbf{y} \right\},$$

with some $\rho > 0$ that trades the relative sparsity of $\mathbf{x}$ and $\mathbf{y}$.

We consider the case where $\rho \equiv 1$ in the following few slides.

ℓ_0-"norm" approach ($\rho \equiv 1$)

Define $\mathbf{A} := \begin{bmatrix} \mathbf{I} & \mathbf{D} \end{bmatrix}$ and $\hat{\mathbf{u}} := \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix}$.

$$\hat{\mathbf{u}} \in \arg\min_{\mathbf{u}\in\mathbb{R}^{2p}} \left\{ \|\mathbf{u}\|_0 : \mathbf{z} = \mathbf{A}\mathbf{u} \right\}.$$

Uncertainty principle

Theorem (Uncertainty principle[1] [9])

For any $\mathbf{x} \in \mathbb{R}^p$ such that $\mathbf{x} \neq \mathbf{0}$, $\|\mathbf{x}\|_0 + \|\mathbf{D}\mathbf{x}\|_0 \geq 2\sqrt{p}$.

Theorem ([8, 12])

If $\left\|\mathbf{x}^\natural\right\|_0 + \left\|\mathbf{y}^\natural\right\|_0 < \sqrt{p}$, then $\hat{\mathbf{u}}$ is uniquely defined and $\hat{\mathbf{u}} = \mathbf{u}^\natural$, or $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{x}^\natural, \mathbf{y}^\natural)$.

Proof.

By definition $\mathrm{null}\,(\mathbf{A}) = \left\{(\mathbf{x}^T, (-\mathbf{D}\mathbf{x})^T)^T : \mathbf{x} \in \mathbb{R}^p\right\}$.

Suppose we have two estimates $\hat{\mathbf{u}}_1 := (\hat{\mathbf{x}}_1^T, \hat{\mathbf{y}}_1^T)^T$ and $\hat{\mathbf{u}}_2 := (\hat{\mathbf{x}}_2^T, \hat{\mathbf{y}}_2^T)^T$ such that $\mathbf{A}\hat{\mathbf{u}}_1 = \mathbf{A}\hat{\mathbf{u}}_2 = \mathbf{z}$. Then $\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2 \in \mathrm{null}\,(\mathbf{A})$ and thus $\hat{\mathbf{x}}_1 - \hat{\mathbf{x}}_2 = -\mathbf{D}\,(\hat{\mathbf{y}}_1 - \hat{\mathbf{y}}_2)$.

By the uncertainty principle we have either $\|\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2\|_0 \geq 2\sqrt{p}$ or $\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2 = \mathbf{0}$. By definition $\|\hat{\mathbf{u}}_1\|_0 < \sqrt{p}$ and $\|\hat{\mathbf{u}}_2\|_0 < \sqrt{p}$, which means that $\|\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2\|_0 < 2\sqrt{p}$. Thus we conclude $\hat{\mathbf{u}}_1 = \hat{\mathbf{u}}_2$. □

[1] Heisenberg's uncertainty principle in quantum mechanics is proved by a continuous counterpart of this uncertainty principle [24]. Indeed, Heisenberg's uncertainty principle, unlike many physics laws, is not concluded from experimental results but is a direct mathematical result.

Generalization via incoherence

Consider the following generalization.

Problem

Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p\times p}$ *be two orthogonal matrices. Let* $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ *be sparse, and define* $\mathbf{u}^\natural := ((\mathbf{x}^\natural)^T, (\mathbf{y}^\natural)^T)^T$. *How do we estimate* $\mathbf{x}^\natural$ *and* $\mathbf{y}^\natural$ *given*

$$\mathbf{z} := \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix} \mathbf{u}^\natural := \mathbf{A}\mathbf{u}^\natural?$$

Can we still solve the problem by the following approach?

ℓ_0-"norm" approach

$$\hat{\mathbf{u}} := \begin{bmatrix} \hat{\mathbf{x}} \\ \hat{\mathbf{y}} \end{bmatrix} := \arg\min_{\mathbf{u}\in\mathbb{R}^{2p}} \left\{ \|\mathbf{u}\|_0 : \mathbf{z} = \mathbf{A}\mathbf{u} \right\}.$$

Incoherence and generalized uncertainty principle

Definition (Incoherence [12, 13])

Two orthogonal matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p \times p}$ are mutually incoherent if with some $K > 0$,

$$\sqrt{p} \max_{1 \leq \ell, k \leq p} \{|\langle \mathbf{u}_\ell, \mathbf{v}_k \rangle|\} \leq K,$$

where $\mathbf{u}_\ell / \mathbf{v}_k$ denotes the ℓth/kth column of $\mathbf{U}/\mathbf{V}$.

Example (A maximally incoherent example)

Take $\mathbf{U} := \mathbf{I}$ and $\mathbf{V} := \mathbf{D}$ the DCT matrix. Then $\mathbf{U}$ and $\mathbf{V}$ are mutually incoherent with $K = 1$, which achieves the lower bound of K.

Theorem (Welch bound [23])

Let $\mathbf{A} := [\mathbf{a}_1, \ldots, \mathbf{a}_{p_2}] \in \mathbb{R}^{p_1 \times p_2}$, $p_1 < p_2$, *such that* $\|\mathbf{a}_j\|_2 = 1$ *for all* $j \in \{1, \ldots, p_2\}$. *Then*

$$\max_{i,j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle| \geq \sqrt{\frac{p_2 - p_1}{p_1 (p_2 - 1)}}.$$

Observation: $K \geq \sqrt{\frac{p}{p-1}}$.

Incoherence and generalized uncertainty principle

Theorem (Generalized uncertainty principle [12, 13])

Let $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{p \times p}$ be mutually incoherent orthogonal matrices with parameter K. Let $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^p$ such that $\mathbf{z} = \mathbf{U}\mathbf{x} = \mathbf{V}\mathbf{y}$. Then

$$\|\mathbf{x}\|_0 + \|\mathbf{y}\|_0 \geq \frac{2\sqrt{p}}{K}.$$

Similarly we can prove the following result.

Theorem ([12, 13])

Assume that $\mathbf{U}, \mathbf{V}$ are mutually incoherent orthogonal matrices with parameter $K > 0$. If $\left\|\mathbf{x}^\natural\right\|_0 + \left\|\mathbf{y}^\natural\right\|_0 < \frac{\sqrt{p}}{K}$, then $\hat{\mathbf{u}}$ is uniquely defined and $\hat{\mathbf{u}} = \mathbf{u}^\natural$, or $(\hat{\mathbf{x}}, \hat{\mathbf{y}}) = (\mathbf{x}^\natural, \mathbf{y}^\natural)$.

Outline

- Today
  1. Source separation problem
  2. Incoherence and uncertainty principle
  3. *General recipe for source separation*
  4. Phase transition via statistical dimension
  5. Phase transition via convex polytopes
  6. Nonsmooth convex minimization by smoothing
- Next week
  1. Constrained convex minimization

Computational issue

Consider the general estimator of $(\mathbf{x}^\natural, \mathbf{y}^\natural)$ given $\mathbf{z} := \mathbf{U}\mathbf{x}^\natural + \mathbf{V}\mathbf{y}^\natural$.

ℓ_0-"norm" approach

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 + \rho \|\mathbf{y}\|_0 : \mathbf{z} = \mathbf{U}\mathbf{x} + \mathbf{V}\mathbf{y} \right\}.$$

with some $\rho > 0$ that trades the relative sparsity of $\mathbf{x}$ and $\mathbf{y}$.

Observation: Since $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{U}\mathbf{x} + \mathbf{V}\mathbf{y}$ is a linear mapping, there exists a matrix $\mathbf{A}$ such that $\mathbf{z} = \mathbf{A}\tilde{\mathbf{x}}^\natural$, where $\tilde{\mathbf{x}}^\natural := ((\mathbf{x}^\natural)^T, (\mathbf{y}^\natural)^T)^T$. In fact $\mathbf{A} := \begin{bmatrix} \mathbf{U} & \mathbf{V} \end{bmatrix}$.

Tractability

Choosing $\rho = 1$, we have

$$\hat{\tilde{\mathbf{x}}} \in \arg\min_{\tilde{\mathbf{x}}\in\mathbb{R}^{2p}} \left\{ \|\tilde{\mathbf{x}}\|_0 : \mathbf{z} = \mathbf{A}\tilde{\mathbf{x}} \right\}.$$

Recall from Lecture 4 that this procedure is *NP-hard*.

Formulation with the ℓ_1-norm

Recall the basis pursuit denoising estimator for compressed sensing.

Definition (Basis pursuit denosing)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural$. The basis pursuit denoising estimator for $\mathbf{x}^\natural$ is given by

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

for some $\kappa \geq 0$.

It is natural to consider the following *convex optimization* analogy with $\kappa = 0$.

ℓ_1-norm approach

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x}, \mathbf{y} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 + \rho \|\mathbf{y}\|_1 : \mathbf{z} = \mathbf{U}\mathbf{x} + \mathbf{V}\mathbf{y} \right\}$$

with some $\rho > 0$.

Generalization: Define atomic sets $\mathcal{A}_{\mathbf{x}}$ as the set of columns of $\mathbf{U}$ and $\mathcal{A}_{\mathbf{y}}$ as the set of columns of $\mathbf{V}$. Let $\tilde{\mathbf{x}}^\natural = \mathbf{U}\mathbf{x}^\natural$ and $\tilde{\mathbf{y}}^\natural = \mathbf{V}\mathbf{y}^\natural$. Then, we equivalently have

$$(\hat{\tilde{\mathbf{x}}}, \hat{\tilde{\mathbf{y}}}) \in \arg\min_{\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathbb{R}^p} \left\{ \|\tilde{\mathbf{x}}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\tilde{\mathbf{y}}\|_{\mathcal{A}_{\mathbf{y}}} : \mathbf{z} = \tilde{\mathbf{x}} + \tilde{\mathbf{y}} \right\}$$

with some $\rho > 0$.

Atomic norms revisited

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.

Definition (Gauge function)

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

$$g_{\mathcal{C}}(\mathbf{x}) := \inf\left\{t : \mathbf{x} = t\mathbf{c} \text{ with some } \mathbf{c} \in \mathcal{C}, t > 0\right\}, \quad \forall \mathbf{x} \in \mathbb{R}^p.$$

Definition (Atomic norm)

Let $\mathcal{A}$ be an *atomic set* in $\mathbb{R}^p$, the **atomic norm** associated with $\mathcal{A}$ 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 $\mathcal{A}$.

General recipe for source separation

Problem

Source separation Let $\mathcal{A}_{\mathbf{x}}$ and $\mathcal{A}_{\mathbf{y}}$ be two atomic sets in $\mathbb{R}^p$, and let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to $\mathcal{A}_{\mathbf{x}}$ and $\mathcal{A}_{\mathbf{y}}$ respectively. How do we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{y}^\natural$?

A general recipe

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} : \mathbf{z} = \mathbf{x} + \mathbf{y} \right\}$$

with some $\rho > 0$. In the sequel, we consider how to choose ρ.

Alternative formulations

Other variants are possible. For instance, consider the following constrained variant

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} : \mathbf{z} = \mathbf{x} + \mathbf{y}, \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} \leq \kappa \right\}.$$

When $\kappa = \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_{\mathbf{y}}}$, the true vectors are feasible. As compared to the regularized version, the difficulty of choosing ρ shifts to the difficulty of choosing κ.

Example: Robust PCA

Problem (Robust principal component analysis (PCA) [3])

Let $\mathbf{X} \in \mathbb{R}^{p \times p}$ be sparse and $\mathbf{Y} \in \mathbb{R}^{p \times p}$ be low-rank. How do we estimate $\mathbf{X}$ and $\mathbf{Y}$ given $\mathbf{Z} := \mathbf{X} + \mathbf{Y}$?

Observation:

- $\mathbf{X}$ is *simple* with respect to the atomic set $\mathcal{A}_{\mathbf{X}} := \left\{ \mathbf{A}_{\mathbf{X}} : \|\mathbf{A}_{\mathbf{X}}\|_0 = 1, \|\mathbf{A}_{\mathbf{X}}\|_F = 1 \right\}$, and
- $\mathbf{Y}$ is *simple* with respect to the atomic set $\mathcal{A}_{\mathbf{Y}} := \left\{ \mathbf{A}_{\mathbf{Y}} : \mathrm{rank}(\mathbf{A}_{\mathbf{Y}}) = 1, \|\mathbf{A}_{\mathbf{Y}}\|_F = 1 \right\}$.

Atomic norm approach

$$(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) \in \arg \min_{\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{p \times p}} \left\{ \|\mathbf{X}\|_{\mathcal{A}_{\mathbf{X}}} + \rho \|\mathbf{Y}\|_{\mathcal{A}_{\mathbf{Y}}} \right\}$$

with some $\rho > 0$. Theory states that $\rho = 1/\sqrt{p}$ is nearly optimal.

Recall that $\|\mathbf{X}\|_{\mathcal{A}_{\mathbf{X}}} = \|\mathrm{vec}(\mathbf{X})\|_1$ and $\|\mathbf{Y}\|_{\mathcal{A}_{\mathbf{Y}}} = \|\mathbf{Y}\|_{S_1}$.

Outline

- Today
  1. Source separation problem
  2. Incoherence and uncertainty principle
  3. General recipe for source separation
  4. *Phase transition via statistical dimension*
  5. Phase transition via convex polytopes
  6. Nonsmooth convex minimization by smoothing
- Next week
  1. Constrained convex minimization

Incoherence revisited

Problem

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. How to we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{y}^\natural$?

Example (A coherent example)

When $\mathcal{A}_\mathbf{x} := \mathcal{A}_\mathbf{y} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$, it is again impossible to recover $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ perfectly.

Example (An incoherent example)

When $\mathcal{A}_\mathbf{x} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$ and $\mathcal{A}_\mathbf{y} := \mathbf{D}\mathcal{A}_\mathbf{x}$ with the DCT matrix $\mathbf{D}$, we obtain the incoherent spikes and sines model.

Random basis model

Now we introduce an orthogonal matrix, or a *change of basis* for one atomic set, to model the incoherence.

Problem

Let $\mathbf{Q} \in \mathbb{R}^{p \times p}$ be an orthogonal matrix. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. How do we estimate $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$?

Example (An incoherent example)

When $\mathcal{A}_\mathbf{x} := \mathcal{A}_\mathbf{y} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_p\}$ and $\mathbf{Q} := \mathbf{D}$ is the DCT matrix, we obtain the solvable spikes and sines model.

Insight: The recovery performance depends on the choice of the matrix $\mathbf{Q}$.

Random basis model

Let $\mathbf{Q} \in \mathbb{R}^{p \times p}$ be a *random orthogonal matrix*. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. What is the probability of perfectly recovering $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$?

Rigorous definition of the *random orthogonal matrix*

Definition (Orthogonal group)

A matrix $\mathbf{Q} \in \mathbb{R}^{p \times p}$ is orthogonal if $\mathbf{Q}^T\mathbf{Q} = \mathbf{Q}\mathbf{Q}^T = \mathbf{I}$.

The set of orthogonal matrices in $\mathbb{R}^{p \times p}$, called the orthogonal group, is denoted by $\mathcal{O}_p$.

Definition (* Haar measure on $\mathcal{O}_p$, cf. [19] for a rigorous definition)

A Haar measure on $\mathcal{O}_p$ is a measure μ on the Borel subsets of $\mathcal{O}_p$ such that for each Borel subset $\mathcal{E}$,

$$\mu(\mathcal{E}) = \mu(\mathbf{Q}\mathcal{E}) := \mu(\{\mathbf{Q}e : e \in \mathcal{E}\}).$$

Insight: The definition is an analogy of the *uniform distribution* for $\mathcal{O}_p$.

Example ([2])

Let $\mathbf{M} \in \mathbb{R}^{p \times p}$ be a matrix of i.i.d. standard Gaussian random variables, and let $\mathbf{M} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$ be its singular value decomposition. Then $\mathbf{U}$ is a random matrix drawn from the Haar measure on $\mathcal{O}_p$.

Definition (Random basis)

A random basis of $\mathbb{R}^p$ is a random matrix drawn from the Haar measure on $\mathcal{O}_p$.

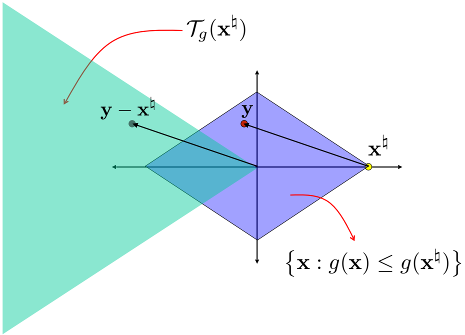
Condition of perfect recovery via tangent cones

Recall the definition of a tangent cone.

Definition (Tangent cone)

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

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



Refined random basis model

Refined random basis model [18]

Let $\mathbf{Q} \in \mathbb{R}^{p\times p}$ be a *random basis*. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. What is the probability of perfectly recovering $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$?

[2] To be defined later. For now, think of them as the Gaussian widths of the cones.

Refined random basis model

Refined random basis model [18]

Let $\mathbf{Q} \in \mathbb{R}^{p\times p}$ be a *random basis*. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. What is the probability of perfectly recovering $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ given $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$?

Define

$$(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) := \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} + \rho\,\|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} : \mathbf{x} + \mathbf{Q}\mathbf{y} = \mathbf{z} \right\}$$

Theorem ([1, 17])

Let $d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)\right)$ and $d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right)$ denote the statistical dimensions[2] of the tangent cones $\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)$ and $\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)$ respectively. Then there exists a $\rho > 0$ such that $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) = (\mathbf{x}^\natural, \mathbf{y}^\natural)$ with probability at least $1-\eta$ if

$$\frac{1}{p}\left[d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right)\right] \leq 1 - \sqrt{\frac{8\log(4/\eta)}{p}}.$$

[2] To be defined later. For now, think of them as the Gaussian widths of the cones.

An equivalent formulation

First we consider an equivalent formulation of

$$(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) := \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho\,\|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} : \mathbf{x} + \mathbf{Q}\mathbf{y} = \mathbf{z} \right\}.$$

Proposition

Let $\mathbf{x}^\natural, \mathbf{y}^\natural \in \mathbb{R}^p$ *and* $\mathbf{Q} \in \mathbb{R}^{p\times p}$ *be given, and let* $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$. *Define*

$$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} : \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} \le \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_{\mathbf{y}}}, \mathbf{x} + \mathbf{Q}\mathbf{y} = \mathbf{z} \right\}.$$

Then there exists a $\rho > 0$ *such that* $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) = (\hat{\mathbf{x}}', \hat{\mathbf{y}}')$.

Proof.

We can use similar arguments as in Lecture 4. □

Condition of perfect recovery via tangent cones

Recall

$$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} : \|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} \le \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_\mathbf{y}}, \mathbf{z} = \mathbf{x} + \mathbf{Q}\mathbf{y} \right\}.$$

Observation 1

$\mathbf{x}^\natural + \mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)$ includes all $\mathbf{x}$ such that $\|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} \le \left\|\mathbf{x}^\natural\right\|_{\mathcal{A}_\mathbf{x}}$.

$\mathbf{x}^\natural + \mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)$ includes *all possible minimizers ignoring the constraint*.

Observation 2

$\mathbf{y}^\natural + \mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)$ includes all $\mathbf{y} \in \mathbb{R}^p$ such that $\|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} \le \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_\mathbf{y}}$.

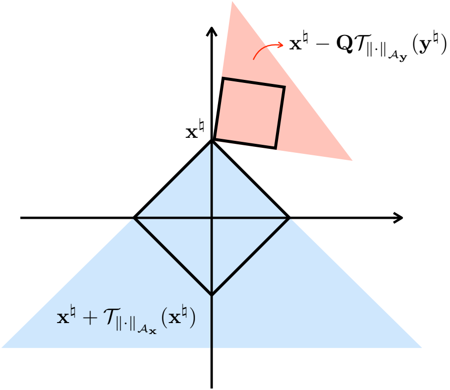
$\mathbf{x}^\natural - \mathbf{Q}\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)$ includes all $\mathbf{x} \in \mathbb{R}^p$ such that $\|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} \le \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_\mathbf{y}}$ and $\mathbf{z} = \mathbf{x} + \mathbf{Q}\mathbf{y}$.

$\mathbf{x}^\natural - \mathbf{Q}\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)$ includes *all feasible points*.

Condition of perfect recovery via tangent cones

Proposition ([1, 18])

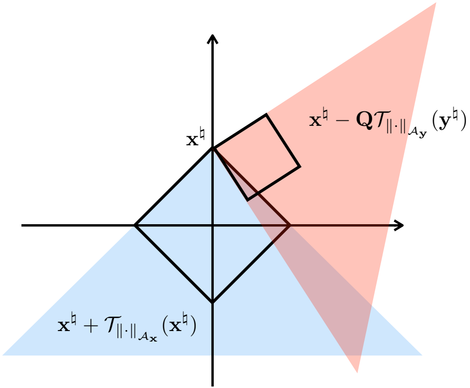
$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)$ *if and only if* $\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right) \cap \left(-\mathbf{Q}\,\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right) = \{\mathbf{0}\}$.



Condition of perfect recovery via tangent cones

Proposition ([1, 18])

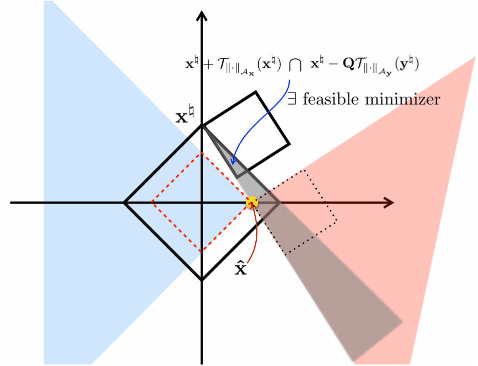
$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)$ *if and only if* $\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right) \cap \left(-\mathbf{Q}\,\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right) = \{\mathbf{0}\}$.



Condition of perfect recovery via tangent cones

Proposition ([1, 18])

$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)$ *if and only if* $\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right) \cap \left(-\mathbf{Q}\,\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right) = \{\mathbf{0}\}$.



Approximate kinematic formula

Definition (Statistical dimension [1])

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

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

where $\Pi_{\mathrm{cl}(\mathcal{C})} : \mathbb{R}^p \to \mathbb{R}^p$ denotes the projection operator onto $\mathrm{cl}(\mathcal{C})$.

Statistical dimension leads to interesting generalizations in the sequel.

Theorem (Approximate kinematic formula [1])

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be convex cones in $\mathbb{R}^p$, and let $\mathbf{Q}$ be a random basis. Then

$$\frac{1}{p}\left[d(\mathcal{C}_1) + d(\mathcal{C}_2)\right] \le 1 - \frac{c_\eta}{\sqrt{p}} \quad \Rightarrow \quad \mathbb{P}\left(\{\mathcal{C}_1 \cap \mathbf{Q}\mathcal{C}_2 = \{\mathbf{0}\}\}\right) \ge 1 - \eta,$$

$$\frac{1}{p}\left[d(\mathcal{C}_1) + d(\mathcal{C}_2)\right] \ge 1 + \frac{c_\eta}{\sqrt{p}} \quad \Rightarrow \quad \mathbb{P}\left(\{\mathcal{C}_1 \cap \mathbf{Q}\mathcal{C}_2 \ne \{\mathbf{0}\}\}\right) \ge 1 - \eta,$$

with any $\eta \in (0,1)$, where $c_\eta := \sqrt{8\log(4/\eta)}$.

Proof: This is an approximation of the kinematic formula from [15].

Performance guarantee

Recall the definition

$$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') := \arg \min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} : \|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} \leq \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_\mathbf{y}}, \mathbf{x} + \mathbf{Q}\mathbf{y} = \mathbf{z} \right\}.$$

Theorem ([1])

Let $\eta \in (0,1)$. If

$$d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right) \leq p - c_\eta \sqrt{p},$$

where $c_\eta := \sqrt{8\log(4/\eta)}$, then $(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)$ with probability at least $1 - \eta$.

Proof.

Combine the condition of perfect recovery and the approximate kinematic formula. Then apply the equivalence relation between $(\hat{\mathbf{x}}', \hat{\mathbf{y}}')$ and $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho))$. □

Performance guarantee

Recall the definition

$$(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) := \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} + \rho \|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} : \mathbf{x} + \mathbf{Q}\mathbf{y} = \mathbf{z} \right\}.$$

Corollary

Let $\eta \in (0,1)$. *If*

$$d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right) \leq p - c_\eta\sqrt{p},$$

where $c_\eta := \sqrt{8\log(4/\eta)}$, *then there exists* $\rho > 0$ *such that* $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) = (\mathbf{x}^\natural, \mathbf{y}^\natural)$.

Proof.

Recall the equivalence relation between $(\hat{\mathbf{x}}', \hat{\mathbf{y}}')$ and $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho))$. □

Successful recovery if $p \gtrsim d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{x}}}\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_\mathbf{y}}}\left(\mathbf{y}^\natural\right)\right)$.

Properties of the statistical dimension

Recall the definition of the statistical dimension.

Definition (Statistical dimension [1])

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

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

where $\Pi_{\mathrm{cl}(\mathcal{C})} : \mathbb{R}^p \to \mathbb{R}^p$ denotes the projection operator onto $\mathrm{cl}(\mathcal{C})$.

Proposition ([1, 4])

1. ***(Rotational invariance)*** *Let $\mathcal{C}$ be a convex cone. Then $d(\mathcal{C}) = d(\mathbf{Q}\mathcal{C})$ for any orthogonal matrix $\mathbf{Q}$.*
2. ***(Monotonicity)*** *Let $\mathcal{C}_1 \subseteq \mathcal{C}_2$ be two convex cones. Then $d(\mathcal{C}_1) \leq d(\mathcal{C}_2)$.*
3. ***(Subspace)*** *For each subspace $\mathcal{L} \subseteq \mathbb{R}^p$, $d(\mathcal{L}) = \dim(\mathcal{L})$.*
4. ***(Complementarity)*** *Let $\mathcal{C} \subseteq$ be a convex cone and $\mathcal{C}^\circ$ be its polar cone. Then $d(\mathcal{C}_1) + d(\mathcal{C}^\circ) = p$.*

Observation: Statistical dimension extends the idea of the affine dimension of vector spaces to convex cones.

Some examples

Example (Convex cones [1])

1. Let $\mathcal{C} := \left\{ \mathbf{x} := (x_1, \ldots, x_p)^T : x_i \geq 0 \, \forall i, \mathbf{x} \in \mathbb{R}^p \right\}$. Then $d\left(\mathcal{C}\right) = \frac{1}{2} d$.
2. Let $\mathcal{C} := \left\{ \mathbf{x} := (\tilde{\mathbf{x}}^T, x_p)^T : \|\tilde{\mathbf{x}}\|_2 \leq x_p, \tilde{\mathbf{x}} \in \mathbb{R}^{p-1}, x_p > 0 \right\}$. Then $d\left(\mathcal{C}\right) = \frac{1}{2} d$.
3. Let $\mathcal{C} := \left\{ \mathbf{X} : \mathbf{X} \succeq \mathbf{0}, \mathbf{X} \in \mathbb{R}^{p \times p} \right\}$. Then $d\left(\mathcal{C}\right) = \frac{1}{4} p(p+1)$.

Example (Tangent cones [1, 4])

1. Let $\mathbf{x} \in \mathbb{R}^p$ be s-sparse, and $f : \mathbf{x} \mapsto \|\mathbf{x}\|_1$. Then $d\left(\mathcal{T}_f\left(\mathbf{x}\right)\right) \leq 2s \log\left(\frac{p}{s}\right) + \frac{5}{4} s$.
2. Let $\mathbf{x} := (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ such that $\left|\left\{ i : |x_i| = \|\mathbf{x}\|_\infty \right\}\right| \leq s$, and $f : \mathbf{x} \mapsto \|\mathbf{x}\|_\infty$. Then $d\left(\mathcal{T}_f\left(\mathbf{x}\right)\right) = p - \frac{1}{2} s$.
3. Let $\mathbf{X} \in \mathbb{R}^{p \times p}$ of rank r, and $f : \mathbf{X} \mapsto \|\mathbf{X}\|_{S_1}$. Then $d\left(\mathcal{T}_f\left(\mathbf{X}\right)\right) \leq 3r(2p - r)$.

Relation between Gaussian width and statistical dimension

An equivalent definition of the statistical dimension is given by the following.

Proposition ([1, 4])

Let $\mathcal{C}$ be a convex cone in $\mathbb{R}^p$. The statistical dimension is given by

$$d(\mathcal{C}) := \mathbb{E}\left[\sup_{\mathbf{x}\in\mathcal{C}\cap\mathcal{B}_p} \langle \mathbf{g}, \mathbf{x}\rangle^2\right],$$

where $\mathcal{B}_p$ denotes the unit ℓ_2-norm ball in $\mathbb{R}^p$, and g is a vector of i.i.d. standard Gaussian random variables.

Note that this definition is very close to the definition of the Gaussian width.

Proposition ([1])

Let $\mathcal{C}$ be a convex cone in $\mathbb{R}^p$, and $\mathcal{S}_p$ be the unit ℓ_2-norm sphere. Then

$$[w(\mathcal{C}\cap\mathcal{S}_p)]^2 \leq d(\mathcal{C}) \leq [w(\mathcal{C}\cap\mathcal{S}_p)]^2 + 1,$$

where $w(\cdot)$ denotes the Gaussian width in Lecture 4.

Insight: $[w(\mathcal{C}\cap\mathcal{S}_p)]^2 \sim d(\mathcal{C})$.

Compressed sensing revisited

Recall the following compressed sensing problem, the basis pursuit denoising estimator $\hat{\mathbf{x}}_{\text{BPDN}}$, and the optimality condition.

Problem (Compressed sensing)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ be simple with respect to an atomic set $\mathcal{A}$, and let $\mathbf{A} \in \mathbb{R}^{n\times p}$ with $p > n$. How do we estimate $\mathbf{x}^\natural$ given $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural$ and $\mathbf{A}$?

Definition (Basis pursuit denoising estimator)

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

Proposition ([5])

Define $f : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$. Then $\hat{\mathbf{x}}_{BPDN}$ is uniquely defined and perfectly recovers $\mathbf{x}^\natural$, i.e., $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$, if and only if

$$\mathcal{T}_f\left(\mathbf{x}^\natural\right) \cap \operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}.$$

Compressed sensing revisited

***Fact**

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ be a random matrix of i.i.d. standard Gaussian random variables with $p > n$. Let $\mathcal{L}$ be a $(p-n)$-dimensional subspace in $\mathbb{R}^p$. Then $\operatorname{null}(\mathbf{A})$ is equivalent to $\mathbf{Q}\mathcal{L}$ almost surely, where $\mathbf{Q} \in \mathbb{R}^p$ denotes the random basis.

Thus the probability that $\mathcal{T}_f\left(\mathbf{x}^\natural\right) \cap \operatorname{null}(\mathbf{A}) = \{\mathbf{0}\}$ is equal to the probability that $\mathcal{T}_f\left(\mathbf{x}^\natural\right) \cap \mathbf{Q}\mathcal{L} = \{\mathbf{0}\}$.

Note that $\mathcal{T}_f\left(\mathbf{x}^\natural\right)$ and $\mathcal{L}$ are two convex cones. Thus we can apply the approximate kinematic formula and obtain the following.

Theorem (Performance guarantee with statistical dimension [1])

Assume that $\mathbf{A} \in \mathbb{R}^{n \times p}$ is a matrix of i.i.d. standard Gaussian random variables with $n < p$. Let $\eta \in (0,1)$. Then $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$ with probability at least $1-\eta$ given that

$$n \geq d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) - c_\eta \sqrt{p},$$

where $f : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$, and $c_\eta := \sqrt{8\log(4/\eta)}$.

Compressed sensing revisited

Recall the result we obtained in Lecture 2.

Theorem (Performance guarantee with Gaussian width [5])

Assume that $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a matrix of i.i.d. standard Gaussian random variables with* $n < p$. *Then* $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^{\natural}$ *with probability at least* $1 - \exp\left\{-\frac{1}{2}\left[\sqrt{n} - w\left(\mathcal{S}_p \cap \mathcal{T}_f\left(\mathbf{x}^{\natural}\right)\right)\right]\right\}$ *given that*

$$n \geq w\left(\mathcal{S}_p \cap \mathcal{T}_f\left(\mathbf{x}^{\natural}\right)\right)^2 + 1,$$

where $f : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$, *and* $\mathcal{S}_p$ *denotes the unit* ℓ_2*-norm sphere.*

Insight: $[w(\mathcal{C} \cap \mathcal{S}_p)]^2 \sim d\left(\mathcal{C}\right)$.

What is the benefit of using the statistical dimension?

Making use of the converse part

Recall the approximate kinematic formula.

Theorem (Approximate kinematic formula [1])

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be convex cones in $\mathbb{R}^p$, and let $\mathbf{Q}$ be a random basis. Then

$$\frac{1}{p}\left[d\left(\mathcal{C}_1\right)+d\left(\mathcal{C}_2\right)\right] \leq 1-\frac{c_\eta}{\sqrt{p}} \quad \Rightarrow \quad \mathbb{P}\left(\left\{\mathcal{C}_1 \cap \mathbf{Q}\mathcal{C}_2=\{\mathbf{0}\}\right\}\right) \geq 1-\eta,$$

$$\frac{1}{p}\left[d\left(\mathcal{C}_1\right)+d\left(\mathcal{C}_2\right)\right] \geq 1+\frac{c_\eta}{\sqrt{p}} \quad \Rightarrow \quad \mathbb{P}\left(\left\{\mathcal{C}_1 \cap \mathbf{Q}\mathcal{C}_2 \neq \{\mathbf{0}\}\right\}\right) \geq 1-\eta,$$

with any $\eta \in (0,1)$, where $c_\eta := \sqrt{8\log(4/\eta)}$.

Insight: When $\frac{1}{p}\left[d\left(\mathcal{C}_1\right)+d\left(\mathcal{C}_2\right)\right] \geq 1+\frac{c_\eta}{\sqrt{p}}$, it is *impossible* to have $\mathbb{P}\left(\left\{\mathcal{C}_1 \cap \mathbf{Q}\mathcal{C}_1=\{\mathbf{0}\}\right\}\right)$ arbitrarily close to 1.

A complete result for source separation

Random basis model [18]

Let $\mathbf{Q} \in \mathbb{R}^{p\times p}$ be a random basis. Let $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ be simple with respect to atomic sets $\mathcal{A}_\mathbf{x}$ and $\mathcal{A}_\mathbf{y}$, respectively. Define $\mathbf{z} := \mathbf{x}^\natural + \mathbf{Q}\mathbf{y}^\natural$ and $(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}} : \|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}} \leq \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_\mathbf{y}}, \mathbf{z} = \mathbf{x} + \mathbf{Q}\mathbf{y} \right\}$. What is the probability of $(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)$?

Theorem ([1])

Let $f : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}_\mathbf{x}}$ *and* $g : \mathbf{y} \mapsto \|\mathbf{y}\|_{\mathcal{A}_\mathbf{y}}$.

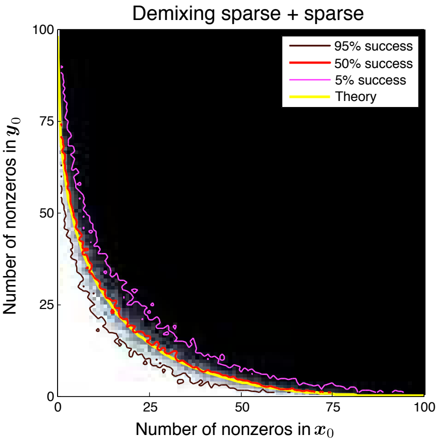
$$
\begin{aligned}
d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_g\left(\mathbf{y}^\natural\right)\right) \leq p - c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left(\left\{(\hat{\mathbf{x}}', \hat{\mathbf{y}}') = (\mathbf{x}^\natural, \mathbf{y}^\natural)\right\}\right) \geq 1 - \eta,\\
d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_g\left(\mathbf{y}^\natural\right)\right) \geq p + c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left(\left\{(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \neq (\mathbf{x}^\natural, \mathbf{y}^\natural)\right\}\right) \geq 1 - \eta,
\end{aligned}
$$

for any $\eta \in (0,1)$, *where* $c_\eta := \sqrt{8\log(4/\eta)}$.

Successful recovery **if and only if** $p \gtrsim d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_g\left(\mathbf{y}^\natural\right)\right)$.

We say there is a *phase transition* at $p \approx d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) + d\left(\mathcal{T}_g\left(\mathbf{y}^\natural\right)\right)$.

Numerical result



A complete result for compressive sensing

Problem (Compressed sensing)

Let $\mathbf{x}^\natural \in \mathbb{R}^p$ *be simple with respect to an atomic set* $\mathcal{A}$*, and let* $\mathbf{A} \in \mathbb{R}^{n\times p}$ *be a matrix of i.i.d. standard Gaussian random variables with* $p > n$*. Define* $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural$ *and* $\hat{\mathbf{x}}_{BPDN} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}} : \mathbf{b} = \mathbf{A}\mathbf{x} \right\}$*. What is the probability of* $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$*?*

Theorem ([1])

Let $f : \mathbf{x} \mapsto \|\mathbf{x}\|_{\mathcal{A}}$*. Then*

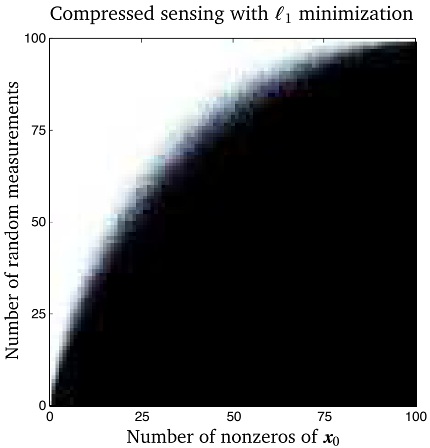
$$
\begin{aligned}
n \geq d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) - c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left(\left\{\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural\right\}\right) \geq 1-\eta,\\
n \leq d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right) + c_\eta\sqrt{p} \quad &\Rightarrow \quad \mathbb{P}\left(\left\{\hat{\mathbf{x}}_{BPDN} \neq \mathbf{x}^\natural)\right\}\right) \geq 1-\eta,
\end{aligned}
$$

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

Successful recovery **if and only if** $n \gtrsim d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right)$.

We say there is a *phase transition* at $n \approx d\left(\mathcal{T}_f\left(\mathbf{x}^\natural\right)\right)$.

Numerical result



Extension to compressive multiple source separation

Problem (Compressive multiple source separation)

Let $\mathbf{A} \in \mathbb{R}^{n \times p}$ with $n < p$. Let $\mathcal{A}_i$, $i = 1, \ldots, N$ be atomic sets in $\mathbb{R}^p$, and $\mathbf{x}_i^\natural \in \mathbb{R}^p$ be simple with respect to $\mathcal{A}_i$ for all $i \in \{1, \ldots, N\}$. Let $\mathbf{Q}_1, \ldots, \mathbf{Q}_N \in \mathbb{R}^{p \times p}$ be independent random bases and define $\mathbf{z} := \mathbf{A}\left(\mathbf{Q}_1\mathbf{x}_1^\natural + \cdots + \mathbf{Q}_N\mathbf{x}_N^\natural\right)$. What is the probability of $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) = (\mathbf{x}_1^\natural, \ldots, \mathbf{x}_N^\natural)$ with

$$(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) \in \arg \min_{\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^p} \Big\{ \|\mathbf{x}_1\|_{\mathcal{A}_1} : \|\mathbf{x}_i\|_{\mathcal{A}_i} \leq \left\|\mathbf{x}_i^\natural\right\|_{\mathcal{A}_i}, i = 2, \ldots, N,$$
$$\mathbf{z} = \mathbf{A}\left(\mathbf{Q}_1\mathbf{x}_1 + \cdot + \mathbf{Q}_N\mathbf{x}_N\right) \Big\}?$$

Extension to compressive multiple source separation

Recall that when we have $\mathbf{z} := \mathbf{x}_1^\natural + \mathbf{x}_2^\natural \in \mathbb{R}^p$, $(\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2) = (\mathbf{x}_1^\natural, \mathbf{x}_2^\natural)$ with high probability if and only if

$$d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_1}}\left(\mathbf{x}_1^\natural\right)\right) + d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_2}}\left(\mathbf{x}_2^\natural\right)\right) \lesssim p = \dim(\mathbf{z}).$$

A reasonable guess

$(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) = (\mathbf{x}_1^\natural, \ldots, \mathbf{x}_N^\natural)$ with high probability if and only if

$$\sum_{i=1}^{N} d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_i}}\left(\mathbf{x}_i^\natural\right)\right) \lesssim n = \dim(\mathbf{z}).$$

Optimality condition

Definition (Minkowski sum)

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two sets. The Minkowski sum of $\mathcal{S}_1$ and $\mathcal{S}_2$ is given by

$$\mathcal{S}_1 + \mathcal{S}_2 := \{\mathbf{s}_1 + \mathbf{s}_2 : \mathbf{s}_1 \in \mathcal{S}_1, \mathbf{s}_2 \in \mathcal{S}_2\}.$$

Theorem ([16])

Define $\mathcal{C}_i := \mathcal{T}_{\|\cdot\|_{\mathcal{A}_i}}\left(\mathbf{x}^\natural\right)$, $i = 1, \ldots, N$, $\mathcal{C}_{N+1} := \mathrm{null}\,(\mathbf{A})$. *We have* $(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N) = (\mathbf{x}_1^\natural, \ldots, \mathbf{x}_N^\natural)$ *if and only if*

$$\mathcal{C}_i \cap \left(-\sum_{j \neq i} \mathcal{C}_j\right) = \{\mathbf{0}\}$$

for all $i \in \{1, \ldots, N+1\}$.

⋆ Phase transition for compressive multiple source separation

Theorem ([16])

Define

$$d_{\max} := \max_{i \in \{1,\ldots,N\}} \left\{ d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_i}}\left(\mathbf{x}^\natural\right)\right)\right\}$$

$$d_{total} := \sum_{i=1}^{N} d\left(\mathcal{T}_{\|\cdot\|_{\mathcal{A}_i}}\left(\mathbf{x}^\natural\right)\right)$$

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

$$n \geq d_{total} + p\left(c_\eta + \sqrt{2c_\eta} d_{\max}\right) \quad \Rightarrow \quad \mathbb{P}\left((\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N) = (\mathbf{x}_1^\natural,\ldots,\mathbf{x}_N^\natural)\right) \geq 1-\eta,$$

$$n \leq d_{total} - p\left(c_\eta + \sqrt{2c_\eta} d_{\max}\right) \quad \Rightarrow \quad \mathbb{P}\left((\hat{\mathbf{x}}_1,\ldots,\hat{\mathbf{x}}_N) \neq (\mathbf{x}_1^\natural,\ldots,\mathbf{x}_N^\natural)\right) \geq 1-\eta,$$

where $c_\eta := \log(4p/\eta)$.

Successful recovery if and only if $n \gtrsim d_{\text{total}}$.

We say there is a *phase transition* at $n \approx d_{\text{total}}$.

Outline

- Today
    1. Source separation problem
    2. Incoherence and uncertainty principle
    3. Phase transition via statistical dimension
    4. *Phase transition via convex polytopes*
    5. Nonsmooth convex minimization by smoothing
- Next week
    1. Constrained convex minimization

Basic notions about convex polytopes

Definition (Convex polytope)

A convex polytope in $\mathbb{R}^n$ is the convex hull of a finite set of points in $\mathbb{R}^n$.

By definition we find the relation between convex polytopes and unit atomic norm balls.

Proposition

A set $\mathcal{P} \subset \mathbb{R}^n$ is a convex polytope if and only if it is a unit atomic norm ball of a finite atomic set in $\mathbb{R}^n$.

Example

Define $\mathbf{e}_i := (\delta_{1,i}, \ldots, \delta n, i)^T \in \mathbb{R}^n$.

Let $\mathcal{A} := \{\mathbf{e}_1, \ldots, \mathbf{e}_n\} \subset \mathbb{R}^n$. Then the unit atomic norm ball associated with $\mathcal{A}$ is a convex polytope called the *simplex*.

Let $\mathcal{A} := \{\pm\mathbf{e}_1, \ldots, \pm\mathbf{e}_n\} \subset \mathbb{R}^n$. The the unit atomic norm ball associated with $\mathcal{A}$ is a convex polytope called the *cross-polytope*.

Basic notions about convex polytopes

Definition (s-face)

An s-face of a convex polytope $\mathcal{P}$ is an s-dimensional face of $\mathcal{P}$.

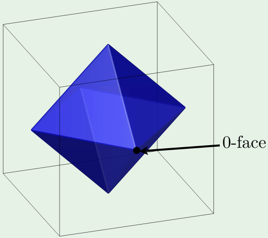
The set of all s-faces of $\mathcal{P}$ is denoted by $\mathcal{F}_s(\mathcal{P})$.

Example

A 0–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a vertex of $\mathcal{P}$.

An $n-1$–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a facet of $\mathcal{P}$.

Example (Cross-polytope)



Basic notions about convex polytopes

Definition (s-face)

An s-face of a convex polytope $\mathcal{P}$ is an s-dimensional face of $\mathcal{P}$.

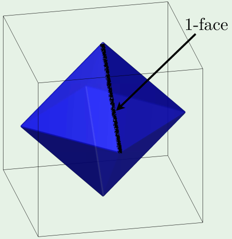
The set of all s-faces of $\mathcal{P}$ is denoted by $\mathcal{F}_s(\mathcal{P})$.

Example

A 0–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a vertex of $\mathcal{P}$.

An $n-1$–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a facet of $\mathcal{P}$.

Example (Cross-polytope)



Basic notions about convex polytopes

Definition (s-face)

An s-face of a convex polytope $\mathcal{P}$ is an s-dimensional face of $\mathcal{P}$.

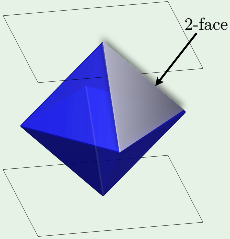
The set of all s-faces of $\mathcal{P}$ is denoted by $\mathcal{F}_s(\mathcal{P})$.

Example

A 0–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a vertex of $\mathcal{P}$.

An $n-1$–face of a convex polytope $\mathcal{P} \subset \mathbb{R}^n$ is a facet of $\mathcal{P}$.

Example (Cross-polytope)



Basic notions about convex polytopes

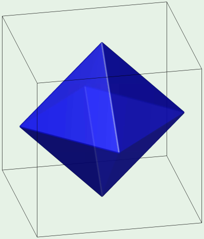
Definition (Centrally symmetric sets)

A pair $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^n$ is called an *antipodal pair* if $\mathbf{x} = -\mathbf{y}$.

A set $\mathcal{E}$ is *centrally symmetric* if for any antipodal pair $(\mathbf{x}, \mathbf{y})$ such that $\mathbf{x} \in \mathcal{E}$, $\mathbf{y} \in \mathcal{E}$.

Example (Cross-polytope)

The cross-polytope (or ℓ_1-ball) $\mathcal{C}$ is *centrally symmetric* since $\forall \mathbf{x} \in \mathcal{C}$, i.e., $\|\mathbf{x}\|_1 \leq 1$, then $\mathbf{y} = -\mathbf{x}$, satisfies $\|\mathbf{y}\|_1 = \|\mathbf{x}\|_1 \leq 1$, so $\mathbf{y} \in \mathcal{C}$.



Basic notions about convex polytopes

Definition (s-neighborliness)

A centrally symmetric convex polytope $\mathcal{P}$ is *s-neighborly* if any $(s+1)$ vertices not including an antipodal pair span a face of $\mathcal{P}$.

Example (Cross-polytope)

The cross-polytope is *2-neighborly*, since any combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$.

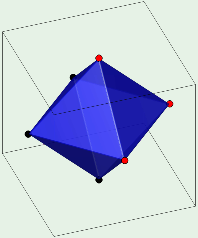


Figure: Combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$

Basic notions about convex polytopes

Definition (s-neighborliness)

A centrally symmetric convex polytope $\mathcal{P}$ is *s-neighborly* if any $(s+1)$ vertices not including an antipodal pair span a face of $\mathcal{P}$.

Example (Cross-polytope)

The cross-polytope is *2-neighborly*, since any combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$.

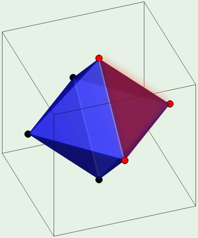


Figure: Combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$

Basic notions about convex polytopes

Definition (s-neighborliness)

A centrally symmetric convex polytope $\mathcal{P}$ is *s-neighborly* if any $(s+1)$ vertices not including an antipodal pair span a face of $\mathcal{P}$.

Example (Cross-polytope)

The cross-polytope is *2-neighborly*, since any combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$.

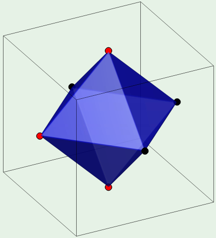


Figure: Combination of 3 vertices, including an antipodal pair, does not span a face of $\mathcal{C}$

Basic notions about convex polytopes

Definition (s-neighborliness)

A centrally symmetric convex polytope $\mathcal{P}$ is *s-neighborly* if any $(s+1)$ vertices not including an antipodal pair span a face of $\mathcal{P}$.

Example (Cross-polytope)

The cross-polytope is *2-neighborly*, since any combination of 3 vertices, not including an antipodal pair, span a face of $\mathcal{C}$.

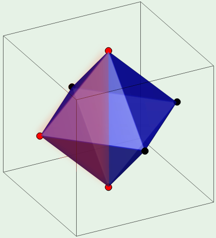


Figure: Combination of 3 vertices, including an antipodal pair, does not span a face of $\mathcal{C}$

An equivalence relation

Consider estimating $\mathbf{x}^\natural \in \mathbb{R}^p$ given $\mathbf{A} \in \mathbb{R}^{n\times p}$, $n < p$, and $\mathbf{b} := \mathbf{A}\mathbf{x}^\natural \in \mathbb{R}^n$ by

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

Denote by $\mathcal{C}$ the cross-polytope in $\mathbb{R}^p$, and define $\mathcal{P} := \mathbf{A}\mathcal{C} := \{\mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathcal{C}\}$. Note that $\mathcal{P}$ is also a convex polytope.

Theorem (ℓ_0/ℓ_1 equivalence [7])

The following two statements are equivalent.

1. *$\mathcal{P}$ has $2p$ vertices and is s-neighborly.*
2. *For every s-sparse $\mathbf{x}^\natural \in \mathbb{R}^p$, $\hat{\mathbf{x}}_{BPDN}$ is uniquely defined and $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$.*

Geometric intuition behind the ℓ_0/ℓ_1 equivalence

Insight 1

A sparse vector $\mathbf{x}^\natural$ is on a k-face of the crosspolytope with $k = \left\|\mathbf{x}^\natural\right\|_0 - 1$.

Insight 2

Let $\mathcal{C} \subset \mathbb{R}^p$ be the crosspolytope and $\mathbf{A} \in \mathbb{R}^{n \times p}$ with $n < p$. Define $\mathcal{P} := \mathbf{A}\mathcal{C}$. Then $\mathcal{F}_\ell(\mathbf{A}\mathcal{C}) \subseteq \mathbf{A}\mathcal{F}_\ell(\mathcal{C})$ for all ℓ.

Some faces of $\mathcal{C}$ *may not survive* after being transformed by $\mathbf{A}$.

Insight 3

Assume $\left\|\mathbf{x}^\natural\right\|_1 = 1$ without loss of generality. To have $\hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural$, it is necessary that $\mathbf{A}\mathbf{x}^\natural$ is on a face of $\mathcal{P} := \mathbf{A}\mathcal{C}$.

Conclusion

It is necessary that all ℓ-faces of $\mathcal{C}$, $0 \leq \ell \leq s - 1$, survive to have $\hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural$ for all $\mathbf{x}^\natural$ being s-sparse.

Geometric intuition behind the ℓ_0/ℓ_1 equivalence

Recall the theorem statement.

Theorem (ℓ_0/ℓ_1 equivalence [7])

The following two statements are equivalent.

1. *$\mathcal{P}$ has $2p$ vertices and is s-neighborly.*
2. *For every s-sparse $\mathbf{x}^\natural \in \mathbb{R}^p$, $\hat{\mathbf{x}}_{BPDN}$ is uniquely defined and $\hat{\mathbf{x}}_{BPDN} = \mathbf{x}^\natural$.*

The conclusion in the previous slide is in fact both necessary and sufficient.

Lemma ([7])

$\mathcal{P} := \mathbf{A}\mathcal{C}$ has $2p$ vertices and is s-neighborly if and only if for all $0 \leq \ell \leq s-1$, $\mathbf{A}\mathcal{F} \in \mathcal{F}_\ell(\mathbf{A}\mathcal{C})$.

Conclusion

$\hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural$ for all $\mathbf{x}^\natural$ being s-sparse, if and only if all ℓ-faces of $\mathcal{C}$, $0 \leq \ell \leq s-1$, survive after being transformed by $\mathbf{A}$.

Face counting

Consider the ratio

$$\gamma_\ell := \frac{|\mathcal{F}_\ell(\mathbf{A}\mathcal{C})|}{|\mathcal{F}_\ell(\mathcal{C})|}.$$

If $\gamma_\ell = 1$ for all $1 \leq \ell \leq s-1$, then $\hat{\mathbf{x}}_{\text{BPDN}} = \mathbf{x}^\natural$ for all s-sparse $\mathbf{x}^\natural$.

Theorem ([6, 10])

Let $\mathbf{A} \in \mathbb{R}^{n\times p}$ be a matrix of i.i.d. standard Gaussian random variables. Consider the triple (n, p, s) with $n = \delta p$ and $s = \rho n$, $0 < \delta, \rho < 1$. Then there exists a function $\rho(\delta)$ such that

$$\lim_{p\to\infty} \gamma_s = \begin{cases} 1 & \rho < \rho(\delta), \\ 0 & \rho > \rho(\delta). \end{cases}$$

Outline

- Today
  1. Source separation problem
  2. Incoherence and uncertainty principle
  3. General recipe for source separation
  4. Phase transition via statistical dimension
  5. Phase transition via convex polytopes
  6. *Selection of the parameter*
  7. Nonsmooth convex minimization by smoothing
- Next week
  1. Constrained convex minimization

Caveat Emptor

The theories presented are based on the equivalence relation between

$$(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} : \mathbf{z} = \mathbf{x} + \mathbf{y} \right\}$$

and

$$(\hat{\mathbf{x}}', \hat{\mathbf{y}}') \in \arg\min_{\mathbf{x},\mathbf{y}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} : \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} \leq \left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_{\mathbf{y}}}, \mathbf{z} = \mathbf{x} + \mathbf{y} \right\}.$$

Caveat Emptor

We select ρ such that $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) = (\hat{\mathbf{x}}', \hat{\mathbf{y}}')$. That is, the selection of ρ requires the information of $\mathbf{y}^\natural$, which is *intractable*.

We show a *semi-practical* approach for a slightly different problem setting.

Problem setting

Corrupted compressive sensing [14]

Let $\mathcal{A}_{\mathbf{x}} \subset \mathbb{R}^p$ and $\mathcal{A}_{\mathbf{y}} \subset \mathbb{R}^n$ be two atomic sets, and $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^n$ be simple with respect to $\mathcal{A}_{\mathbf{x}}$ and $\mathcal{A}_{\mathbf{y}}$ respectively. Let $\mathbf{A} \in \mathbb{R}^{n \times p}$, $n < p$, be a random matrix with i.i.d. Gaussian random variables $\sim \mathcal{N}(0, 1/n)$. Let $\mathbf{z} := \mathbf{A}\mathbf{x}^\natural + \mathbf{y}^\natural + \mathbf{w}$, where $\mathbf{w}$ denotes some unknown noise.

Define the estimator

$$(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho)) \in \arg\min_{\mathbf{x},\mathbf{y} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} : \|\mathbf{z} - (\mathbf{A}\mathbf{x} + \mathbf{y})\|_2 \leq \kappa \right\}.$$

How good is the estimation performance of $(\hat{\mathbf{x}}(\rho), \hat{\mathbf{y}}(\rho))$?

A general bound for arbitrary ρ

Theorem (* Recovery error bound [14])

For any $t_{\mathbf{x}}, t_{\mathbf{y}} > 0$ *such that* $\rho = t_{\mathbf{x}}/t_{\mathbf{y}}$,

$$\sqrt{\left\|\hat{\mathbf{x}}(\rho) - \mathbf{x}^\natural\right\|^2 + \left\|\hat{\mathbf{y}}(\rho) - \mathbf{y}^\natural\right\|^2} \leq \frac{2\kappa}{\epsilon}$$

with probability at least $1 - \exp\left[-(1/2)\left(a_n - \tau - \epsilon\sqrt{n}\right)^2\right]$ *given that* $a_n - \epsilon\sqrt{n} > \tau$, *where*

$$\tau := 2\eta\left(t_{\mathbf{x}}\,\partial\left\|\mathbf{x}^\natural\right\|_{\mathcal{A}_{\mathbf{x}}}\right) + \eta\left(t_{\mathbf{y}}\,\partial\left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_{\mathbf{y}}}\right) + 3\sqrt{2}\pi + \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2\pi}},$$

and $a_n := \mathbb{E}\left[\|\mathbf{g}\|_2\right] \approx \sqrt{n}$, $\mathbf{g} \in \mathbb{R}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The function η is called the *Gaussian distance*, which characterizes how large a set is.

Definition (Gaussian distance [14])

Let $\mathcal{C} \subset \mathbb{R}^n$ *and* $\mathbf{g} \in \mathbb{R}^n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. *The Gaussian distance of* $\mathcal{C}$ *is given by*

$$\eta(\mathcal{C}) := \sqrt{\mathbb{E}\left[\inf_{\mathbf{x}\in\mathcal{C}} \|\mathbf{g} - \mathbf{x}\|_2^2\right]}.$$

Some known upper bounds on the Gaussian distance

Example (ℓ_1-norm)

Let $\mathbf{x} \in \mathbb{R}^p$ be s-sparse. Then $\eta^2(t\,\partial\,\|\mathbf{x}\|_1) \leq 2s\log(p/s) + (3/2)s$ when $t := \sqrt{2\log(p/s)}$.

The following alternative bound is tighter when s/p is large.

Example (ℓ_1-norm)

Let $\mathbf{x} \in \mathbb{R}^p$ be s-sparse. Then $\eta^2(t\,\partial\,\|\mathbf{x}\|_1) \leq p\left[1 - \frac{2}{\pi}\left(1 - \frac{s}{p}\right)^2\right]$ when $t := \sqrt{\frac{2}{\pi}}\left(1 - \frac{s}{p}\right)$.

Example (Nuclear norm)

Let $\mathbf{X} \in \mathbb{R}^{p\times p}$ be rank-r. Then $\eta^2(t\,\partial\,\|\mathbf{X}\|_*) \leq p^2\left[1 - \left(\frac{4}{27}\right)^2\left(1 - \frac{r}{p}\right)^3\right]$ when $t := \frac{4}{27}(p-r)\frac{\sqrt{p-r}}{p}$.

Semi-practical approach

Recall that $t_{\mathbf{x}}$ and $t_{\mathbf{y}}$ are only involved in the definition of τ in the recovery error bound, which establishes a lower bound on the *minimum number of samples* n.

Semi-practical approach [14]

Choose $\rho := \frac{t_{\mathbf{x}}}{t_{\mathbf{y}}}$ to achieve the sharpest theoretical upper bounds on $\eta\left(t_{\mathbf{x}}\,\partial\left\|\mathbf{x}^\natural\right\|_{\mathcal{A}_{\mathbf{x}}}\right)$ and $\eta\left(t_{\mathbf{y}}\,\partial\left\|\mathbf{y}^\natural\right\|_{\mathcal{A}_{\mathbf{y}}}\right)$ (cf. the previous slide).

Warning!

Some knowledge on $\mathbf{x}^\natural$ and $\mathbf{y}^\natural$ is still required. For example, $s := \left\|\mathbf{x}^\natural\right\|_0$ is required for $\|\cdot\|_{\mathcal{A}_{\mathbf{x}}}$ being the ℓ_1-norm, and $r := \mathrm{rank}\left(\mathbf{X}^\natural\right)$ is required for $\|\cdot\|_{\mathcal{A}_{\mathbf{x}}}$ being the nuclear norm.

Outline

- Today
  1. Source separation problem
  2. Incoherence and uncertainty principle
  3. General recipe for source separation
  4. Phase transition via statistical dimension
  5. Phase transition via convex polytopes
  6. Selection of the parameter
  7. *Nonsmooth convex minimization by smoothing*
- Next week
  1. Constrained convex minimization

Composite convex minimization formulation

Problem (Source separation)

Let $\mathcal{A}_{\mathbf{x}}$ and $\mathcal{A}_{\mathbf{y}}$ be two atomic sets in $\mathbb{R}^p$ and $\mathbf{x}^\natural \in \mathbb{R}^p$ and $\mathbf{y}^\natural \in \mathbb{R}^p$ are simple with respect to $\mathcal{A}_{\mathbf{x}}$ and $\mathcal{A}_{\mathbf{y}}$ respectively. Let $\mathbf{z} := \mathbf{x}^\natural + \mathbf{y}^\natural$. We consider the estimator

$$(\hat{\mathbf{x}}, \hat{\mathbf{y}}) \in \arg\min_{\mathbf{x},\mathbf{y} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\mathbf{y}\|_{\mathcal{A}_{\mathbf{y}}} : \mathbf{z} = \mathbf{x} + \mathbf{y} \right\}.$$

Equivalent composite convex minimization formulation

$$\begin{cases} \hat{\mathbf{x}} \in \arg\min_{\mathbf{x} \in \mathbb{R}^p} \left\{ \|\mathbf{x}\|_{\mathcal{A}_{\mathbf{x}}} + \rho \|\mathbf{z} - \mathbf{x}\|_{\mathcal{A}_{\mathbf{y}}} \right\} \\ \hat{\mathbf{y}} := \mathbf{z} - \hat{\mathbf{x}} \text{ (trivial)} \end{cases}.$$

- If $\|\cdot\|_{\mathcal{A}_{\mathbf{x}}}$ or $\|\cdot\|_{\mathcal{A}_{\mathbf{y}}}$ is smooth, we can apply algorithms such as ISTA or FISTA.
- What can we do if both $\|\cdot\|_{\mathcal{A}_{\mathbf{x}}}$ and $\|\cdot\|_{\mathcal{A}_{\mathbf{y}}}$ are *nonsmooth*?

Smoothing for nonsmooth composite convex minimization

Now we consider the general nonsmooth convex minimization problem.

Problem (Nonsmooth composite convex minimization)

$$\boxed{F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \{F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x})\}} \tag{1}$$

where f and g are both proper, closed, convex and nonsmooth.

Smoothing approach

Approximate f by a *smooth* function $\tilde{f}$. Then, use the following approximation

$$\tilde{F}^* := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \tilde{F}(\mathbf{x}) := \tilde{f}(\mathbf{x}) + g(\mathbf{x}) \right\}$$

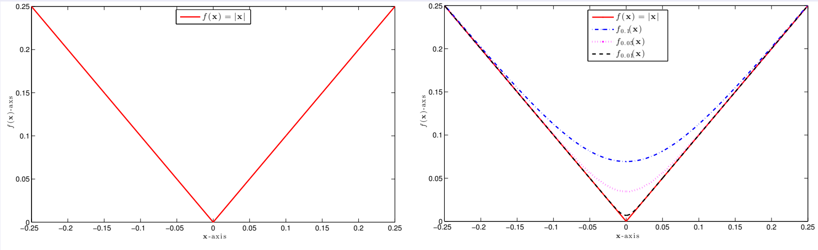
and obtain a numerical solution by the composite minimization algorithms, such as ISTA or FISTA.

Terminology

$\tilde{f}$ is called a *smoother* of f.

Illustration of the smoothing idea

Example: 1-dimensional function $f(\mathbf{x}) = |\mathbf{x}|$ and its smoother



Here, $f_\gamma(\mathbf{x}) = \gamma \log(e^{\mathbf{x}/\gamma} + e^{-\mathbf{x}/\gamma})$ is a smoother of $f(\mathbf{x}) = |\mathbf{x}|$.

Example (Multidimensional case)

$f_\gamma(\mathbf{x}) := \gamma \sum_{i=1}^{n} \log\left[\exp((\mathbf{A}\mathbf{x} - \mathbf{b})_i/\gamma) + \exp(-(\mathbf{A}\mathbf{x} - \mathbf{b})_i/\gamma)\right]$ is a smoother of $f(\mathbf{x}) := \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$.

Smoothable functions

Definition (Smoothable function)

$f \in \mathcal{F}(\mathbb{R}^p)$ is called smoothable over a convex set $\mathcal{X}$ if:

1. There exists $(\gamma, D_{\mathcal{X}}, L) \in \mathbb{R}^3_{++}$ and $f_\gamma \in \mathcal{F}_L^{1,1}(\mathcal{X})$ such that

$$f_\gamma(\mathbf{x}) - \gamma D_{\mathcal{X}} \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x}) + \gamma D_{\mathcal{X}}, \quad \forall \mathbf{x} \in \mathcal{X}, \tag{2}$$

2. f_γ is convex and its gradient is Lipschitz continuous with constant L_γ over $\mathcal{X}$, i.e.:

$$\|\nabla f_\gamma(\mathbf{x}) - \nabla f(\hat{\mathbf{x}})\|^* \leq L_\gamma \|\mathbf{x} - \hat{\mathbf{x}}\|, \quad \mathbf{x}, \hat{\mathbf{x}} \in \mathcal{X}.$$

Smoothable functions

One strategy

- Smooth f by $f_\gamma \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$.
- Solve the smoothed problem

$$F_\gamma^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F_\gamma(\mathbf{x}) := f_\gamma(\mathbf{x}) + g(\mathbf{x}) \right\}. \tag{3}$$

  by **FISTA** to obtain a solution $\mathbf{x}_\gamma^\star$.
- Characterize how $\mathbf{x}_\gamma^\star$ approximates a true solution $\mathbf{x}^\star$ of (1).

Then using [fast] gradient algorithms for the smoothed problem. [3]

[3] When $f \in \mathcal{F}_L^{1,1}(\mathbb{R}^p)$ and g is smoothable, one can smooth g and simply apply the fast gradient method in Lecture 3

Example 1: ℓ_1-norm

Smoothed function f_γ of the ℓ_1-norm $f(\mathbf{x}) := \|\mathbf{x}\|_1$

$$f_\gamma(\mathbf{x}) := \gamma \sum_{i=1}^{p} \log(e^{x_i/\gamma} + e^{-x_i/\gamma}).$$

- f_γ is smooth and ∇f_γ is Lipschitz continuous with $L_{f_\gamma} := 1/\gamma$.
- $f_\gamma(\mathbf{x}) - \gamma p \ln(2) \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$.

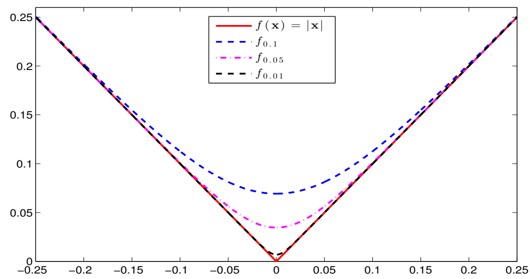
Example 1: ℓ_1-norm

Smoothed function f_γ of the ℓ_1-norm $f(\mathbf{x}) := \|\mathbf{x}\|_1$

$$f_\gamma(\mathbf{x}) := \gamma \sum_{i=1}^{p} \log(e^{x_i/\gamma} + e^{-x_i/\gamma}).$$

- f_γ is smooth and ∇f_γ is Lipschitz continuous with $L_{f_\gamma} := 1/\gamma$.
- $f_\gamma(\mathbf{x}) - \gamma p \ln(2) \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^p$.

1-dimensional function



Example 2: Spectral norm $\lambda_1(\mathbf{X})$

Smoothed function of the spectral norm $f(\mathbf{X}) := \lambda_1(\mathbf{X})$

- The spectral function $f(\mathbf{X}) := \lambda_1(\mathbf{X})$ is the maximum eigenvalue of a symmetric matrix $\mathbf{X} \in \mathbb{S}^{p \times p}$.
- Multinomial logistic smoother $f_\gamma(\mathbf{X})$:

$$f_\gamma(\mathbf{X}) := \gamma \ln\left(\sum_{i=1}^{p} e^{\lambda_i(\mathbf{X})/\gamma}\right).$$

  - f_γ is smooth and ∇f_γ is Lipschitz continuous with $L_{f_\gamma} = \gamma^{-1}$.
  - $f_\gamma(\mathbf{x}) - \gamma \ln(p) \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x})$ for all $\mathbf{X} \in \mathbb{S}^p$.

2-dimensional example

The spectral function $f : \mathbb{S}^2 \to \mathbb{R}$ defined as

$$f(\mathbf{X}) \equiv f\left(\begin{bmatrix} \mathbf{X}_{11} & \mathbf{X}_{12} \\ \mathbf{X}_{12} & \mathbf{X}_{22} \end{bmatrix}\right) := \frac{(\mathbf{X}_{11} + \mathbf{X}_{22})}{2} + \sqrt{\frac{(\mathbf{X}_{11} + \mathbf{X}_{22})^2}{4} - (\mathbf{X}_{11}\mathbf{X}_{22} - \mathbf{X}_{12}^2)}.$$

Proximity functions

Definition (Proximity functions)

A μ_b-strongly convex and continuous function $b_{\mathcal{X}}$ is called a **proximity function** (or prox-function) of a convex set $\mathcal{X}$ if $\mathcal{X} \subseteq \mathrm{dom}(b_{\mathcal{X}})$.

Proximity functions

Definition (Proximity functions)

A μ_b-strongly convex and continuous function $b_{\mathcal{X}}$ is called a **proximity function** (or prox-function) of a convex set $\mathcal{X}$ if $\mathcal{X} \subseteq \text{dom}(b_{\mathcal{X}})$.

Example (Well-known prox-functions)

- $b_{\mathcal{X}}(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|_2^2$ is a prox-function of $\mathcal{X} \equiv \mathbb{R}^p$ (simplest one, $\mu_b = 1$).
- $b_{\mathcal{X}}(\mathbf{x}) := p + \sum_{i=1}^p \mathbf{x}_i \log(\mathbf{x}_i)$ is a prox-function of the standard simplex

$$\mathcal{X} := \{\mathbf{x} \in \mathbb{R}_+^p \ : \ \sum_{i=1}^p \mathbf{x}_i = 1\},$$

where $\mu_b = 1$ measured in ℓ_1-norm (entropy prox-function).

Prox-center and prox-diameter

Definition (Prox-center and prox-diameter)

- A point $\mathbf{x}_c$ defined as
$$\mathbf{x}_c := \arg\min_{\mathbf{x}\in\mathcal{X}} b_{\mathcal{X}}(\mathbf{x})$$
is called the prox-center of $\mathcal{X}$ w.r.t. $b_{\mathcal{X}}$.
- The quantity
$$D^b_{\mathcal{X}} := \sup_{\mathbf{x}\in\mathcal{X}} b_{\mathcal{X}}(\mathbf{x})$$
is called the prox-diameter of $\mathcal{X}$ w.r.t. $b_{\mathcal{X}}$.

Note:

- The point $\mathbf{x}_c$ always exists.
- **Convention:** $b_{\mathcal{X}}(\mathbf{x}_c) = 0$.
- If $\mathcal{X}$ is bounded, then $0 \leq D^b_{\mathcal{X}} < +\infty$.

Example

Example (Entropy function)

- The center point of the entropy prox-function $b_{\mathcal{X}}(\mathbf{x}) := p + \sum_{i=1}^{p} x_i \log(x_i)$ is

$$\mathbf{x}_c := (1/p, 1/p, \cdots, 1/p)^T \in \mathbb{R}^p.$$

- The prox-diameter of $b_{\mathcal{X}}(\mathbf{x}) := p + \sum_{i=1}^{p} x_i \log(x_i)$ is

$$D_{\mathcal{X}}^b := 1 - 1/p.$$

Nesterov's smoothing technique

Problem (Max-structure function)

Given $\mathbf{A} \in \mathbb{R}^{p \times q}$, a convex function $f^* \in \mathcal{F}(\mathbb{R}^q)$ and a nonempty, closed convex set $\mathcal{U} \in \mathbb{R}^q$. Is the following function smoothable?

$$\boxed{f(\mathbf{x}) := \max_{\mathbf{u} \in \mathcal{U}} \{\mathbf{u}^T \mathbf{A}\mathbf{x} - f^*(\mathbf{u})\}, \quad \forall \mathbf{x} \in \mathbb{R}^p.} \tag{4}$$

Nesterov's smoothing technique

Problem (Max-structure function)

Given $\mathbf{A} \in \mathbb{R}^{p\times q}$, a convex function $f^* \in \mathcal{F}(\mathbb{R}^q)$ and a nonempty, closed convex set $\mathcal{U} \in \mathbb{R}^q$. Is the following function smoothable?

$$\boxed{f(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}}\{\mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u})\}, \quad \forall \mathbf{x} \in \mathbb{R}^p.} \tag{4}$$

Definition (**Nesterov's smoother**)

For f given by (4), the function:

$$\boxed{f_\gamma(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}}\{\mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \gamma b_{\mathcal{U}}(\mathbf{u})\}} \tag{5}$$

is a smoother of f, where $b_{\mathcal{U}}$ is a prox-function of $\mathcal{U}$ and $\gamma > 0$ is a smoothness parameter.

Key estimates

Proposition (Nesterov's lemma [20])

- *The function f defined by* (4) *is a* **smoothable function** *by f_γ defined by* (5).
- ***Parameters:*** $(\gamma, D^b_{\mathcal{U}}, L_{f_\gamma})$, *where $D^b_{\mathcal{U}}$ is the prox-diameter of $\mathcal{U}$ and $L_{f_\gamma} := \frac{\|\mathbf{A}\|^2}{\mu_b}$.*
- ***Approximate bound:***

$$f_\gamma(\mathbf{x}) \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x}) + \gamma D^b_{\mathcal{U}}, \quad \forall \mathbf{x} \in \mathbb{R}^p. \tag{6}$$

Example 1: ℓ_1-norm

Problem (ℓ_1-norm)

Is $f(\mathbf{x}) := \|\mathbf{x}\|_1$ a smoothable function? (in Nesterov's sense).

Smoother for f

$$f_\gamma(\mathbf{x}) := \max_{\mathbf{u}\in\mathbb{R}^p}\{\mathbf{x}^T\mathbf{u} - (\gamma/2)\|\mathbf{u}\|_2^2 \ : \ \|\mathbf{u}\|_\infty \leq 1\}.$$

- f_γ is smooth and ∇f_γ is Lipschitz continuous with $L_{f_\gamma} = \gamma^{-1}$.
- $f_\gamma(\mathbf{x}) \leq f(\mathbf{x}) \leq f_\gamma(\mathbf{x}) + \gamma\sqrt{n}$ for all $\mathbf{x} \in \mathbb{R}^p$.

Example 2: Nuclear norm

Is the nuclear norm smoothable?

Problem: $f(\mathbf{X}) := \|\mathbf{X}\|_\star$ - the **nuclear norm** of matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$.

Example 2: Nuclear norm

Is the **nuclear norm** smoothable?

Problem: $f(\mathbf{X}) := \|\mathbf{X}\|_\star$ - the **nuclear norm** of matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$.

Prox-smoother

$$f_\gamma(\mathbf{X}) := \max_{\mathbf{U} \in \mathbb{R}^{n \times p}} \{\mathrm{tr}(\mathbf{X}\mathbf{U}) - (\gamma/2)\|\mathbf{U}\|_F^2 \; : \; \sigma_1(\mathbf{U}) \leq 1\}.$$

- f_γ is smooth and ∇f_γ is Lipschitz continuous with $L_{f_\gamma} = \gamma^{-1}$.
- $f_\gamma(\mathbf{X}) \leq f(\mathbf{X}) \leq f_\gamma(\mathbf{X}) + \gamma\sqrt{mn}$ for all $\mathbf{X} \in \mathbb{R}^{n \times p}$.

Smoothing to nonsmooth minimization

Problem (Nonsmooth composite formulation)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \right\}. \tag{7}$$

Smoothing to nonsmooth minimization

Problem (Nonsmooth composite formulation)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \right\}. \tag{7}$$

Assumption A.3

$f \in \mathcal{F}(\mathbb{R}^p)$ is **smoothable** and $g \in \mathcal{F}_{\mathrm{prox}}(\mathbb{R}^p)$.

Smoothing to nonsmooth minimization

Problem (Nonsmooth composite formulation)

$$F^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) \right\}. \tag{7}$$

Assumption A.3

$f \in \mathcal{F}(\mathbb{R}^p)$ is **smoothable** and $g \in \mathcal{F}_{\mathrm{prox}}(\mathbb{R}^p)$.

Two-step strategy

1. Smooth f by f_γ to obtain the smoothed problem:

$$F_\gamma^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F_\gamma(\mathbf{x}) := f_\gamma(\mathbf{x}) + g(\mathbf{x}) \right\}. \tag{8}$$

2. Apply FISTA to solve the smoothed problem (8).

Smoothing fast proximal-gradient

Smoothing fast proximal-gradient

1. Give an accuracy $\varepsilon > 0$. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point. Set $\gamma := \frac{\varepsilon}{D_{\mathcal{U}}^p}$.

2. Set $\mathbf{y}^0 := \mathbf{x}^0$ and $t_0 := 1$.

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

$$\begin{cases} \mathbf{x}^{k+1} & := \mathrm{prox}_{\lambda g}\left(\mathbf{y}^k - \lambda \nabla f_\gamma(\mathbf{y}^k)\right), \quad \lambda := 1/L_f, \\ t_{k+1} & := 0.5(1 + \sqrt{4t_k^2 + 1}), \\ \eta_{k+1} & := (t_k - 1)/t_{k+1}, \\ \mathbf{y}^{k+1} & := \mathbf{x}^{k+1} + \eta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k). \end{cases} \tag{9}$$

Smoothing fast proximal-gradient

Smoothing fast proximal-gradient

1. Give an accuracy $\varepsilon > 0$. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point. Set $\gamma := \frac{\varepsilon}{D_{\mathcal{U}}^p}$.

2. Set $\mathbf{y}^0 := \mathbf{x}^0$ and $t_0 := 1$.

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

$$\begin{cases} \mathbf{x}^{k+1} & := \mathrm{prox}_{\lambda g}\left(\mathbf{y}^k - \lambda \nabla f_\gamma(\mathbf{y}^k)\right), \quad \lambda := 1/L_f, \\ t_{k+1} & := 0.5(1 + \sqrt{4t_k^2 + 1}), \\ \eta_{k+1} & := (t_k - 1)/t_{k+1}, \\ \mathbf{y}^{k+1} & := \mathbf{x}^{k+1} + \eta_{k+1}(\mathbf{x}^{k+1} - \mathbf{x}^k). \end{cases} \tag{9}$$

Complexity per iteration

- One gradient $\nabla f_\gamma(\mathbf{y}^k)$
- One prox-operator of g
- 8 arithmetic operations for t_{k+1} and η_{k+1};
- 2 more vector additions and 1 scalar-vector multiplication.

The **cost per iteration** is almost the same as in **proximal-gradient scheme**.

Global complexity

Theorem (Global complexity [20])

The worst-case complexity to reach $F(\mathbf{x}^k) - F^\star \leq \varepsilon$ *is*

$$\mathcal{O}\left(2\sqrt{2}\|\mathbf{A}\|_2 \frac{\sqrt{D_{\mathcal{U}}^p} R_0}{\sqrt{\mu_p}\varepsilon}\right), \tag{10}$$

where $R_0 := \max_{\mathbf{x}^\star \in \mathcal{S}^\star} \|\mathbf{x}^0 - \mathbf{x}^\star\|_2$.

Proof of Global complexity

Sketch of proof.

By using FISTA to (8) and the convergence theorem of FISTA, we have

$$F_\gamma(\mathbf{x}^k) - F_\gamma(\mathbf{x}) \leq \frac{2L_{f_\gamma}}{(k+2)^2}\|\mathbf{x}^0 - \mathbf{x}\|_2^2, \ \forall \mathbf{x} \in \mathbb{R}^n.$$

Using (6), we have $F(\mathbf{x}^k) - F(\mathbf{x}^\star) \leq F_\gamma(\mathbf{x}^k) - F_\gamma(\mathbf{x}^\star) + \gamma D_{\mathcal{U}}^p$. Hence

$$F(\mathbf{x}^k) - F(\mathbf{x}^\star) \leq \frac{2\|\mathbf{A}\|_2^2}{\gamma(k+2)^2}R_0^2 + \gamma D_{\mathcal{U}}^p = \varepsilon.$$

Minimizing the right-hand side $s(\gamma) := \frac{2\|\mathbf{A}\|_2^2}{\gamma(k+2)^2}R_0^2 + \gamma D_{\mathcal{U}}^p$ w.r.t. γ, we have $\gamma = \frac{\sqrt{2}\|\mathbf{A}\|_2 R_0}{(k+2)\sqrt{D_{\mathcal{U}}^p}}$.
Using this γ and the fact $s(\gamma) = \varepsilon$, we $\gamma = \frac{\varepsilon}{D_{\mathcal{U}}^p}$ and

$$k + 2 \geq 2\sqrt{2}\|\mathbf{A}\|_2 \frac{\sqrt{D_{\mathcal{U}}^p}R_0}{\sqrt{\mu_p}\varepsilon},$$

which leads to (10). □

Example: Robust PCA

Problem (**RPCA problem**)

$$F^\star := \min_{\mathbf{L}\in\mathbb{R}^{n\times p}} \left\{ F(\mathbf{L}) := \underbrace{\|\mathrm{vec}(\mathbf{M}-\mathbf{L})\|_1}_{f(\mathbf{L})} + \underbrace{\lambda\|\mathbf{L}\|_*}_{g(\mathbf{L})} \right\}.$$

Example: Robust PCA

Problem (**RPCA problem**)

$$F^\star := \min_{\mathbf{L} \in \mathbb{R}^{n \times p}} \left\{ F(\mathbf{L}) := \underbrace{\|\mathrm{vec}(\mathbf{M} - \mathbf{L})\|_1}_{f(\mathbf{L})} + \underbrace{\lambda \|\mathbf{L}\|_*}_{g(\mathbf{L})} \right\}.$$

Strategy

- **Case 1**: Smooth $f(\mathbf{L}) := \|\mathrm{vec}(\mathbf{M} - \mathbf{L})\|_1$ by

$$f_\gamma(\mathbf{L}) := \gamma \sum_{i,j} \log(e^{(\mathbf{M}_{ij} - \mathbf{L}_{ij})/\gamma} + e^{-(\mathbf{M}_{ij} - \mathbf{L}_{ij})/\gamma}).$$

- **Case 2**: Smooth $g(\mathbf{L}) := \|\mathbf{L}\|_*$ by

$$g_\gamma(\mathbf{L}) := \max_{\mathbf{U}} \left\{ \mathrm{tr}(\mathbf{L}^T \mathbf{U}) - (\gamma/2)\|\mathbf{U}\|_F^2 \mid \lambda_1(\mathbf{U}) \leq 1 \right\}.$$

A self-concordant barrier analogue of the smoothing approach

Problem (Max-structure function)

Given $\mathbf{A} \in \mathbb{R}^{p \times q}$, a convex function $f^* \in \mathcal{F}(\mathbb{R}^q)$ and a nonempty, closed convex set $\mathcal{U} \in \mathbb{R}^q$. Is the following function smoothable?

$$\boxed{f(\mathbf{x}) := \max_{\mathbf{u} \in \mathcal{U}} \{\mathbf{u}^T \mathbf{A}\mathbf{x} - f^*(\mathbf{u})\}, \quad \forall \mathbf{x} \in \mathbb{R}^p.}$$

Definition (**Nesterov's smoother**)

$$\boxed{f_\gamma(\mathbf{x}) := \max_{\mathbf{u} \in \mathcal{U}} \{\mathbf{u}^T \mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \gamma p_{\mathcal{U}}(\mathbf{u})\}}$$

is a smoother of f, where $p_{\mathcal{U}}$ is a prox-function of $\mathcal{U}$ and $\gamma > 0$ is a smoothness parameter.

A self-concordant barrier analogue of the smoothing approach

Problem (Max-structure function)

Given $\mathbf{A} \in \mathbb{R}^{p\times q}$, a convex function $f^* \in \mathcal{F}(\mathbb{R}^q)$ and a nonempty, closed convex set $\mathcal{U} \in \mathbb{R}^q$. Is the following function smoothable?

$$\boxed{f(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}}\{\mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u})\}, \quad \forall \mathbf{x} \in \mathbb{R}^p.}$$

Definition (**Self-concordant barrier smoother** [21])

$$\boxed{f_\sigma(\mathbf{x}) := \max_{\mathbf{u}\in\mathcal{U}}\{\mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \sigma b_{\mathcal{U}}(\mathbf{u})\}}$$

is a smoother of f, where $b_{\mathcal{U}}$ is a self-concordant barrier of $\mathcal{U}$ and $\gamma > 0$ is a smoothness parameter.

Recall: Self-concordant barrier

Definition (Self-concordant function)

A convex function $f : \mathrm{dom}(f) \subset \mathbb{R}^n \to \mathbb{R}$ with an open domain is said to be self-concordant with parameter $M \geq 0$, if $|\phi'''(t)| \leq M\,[\phi''(t)]^{3/2}$, where $\phi(t) := f(\mathbf{x} + t\mathbf{v})$ for all $t \in \mathbb{R}$, $\mathbf{x} \in \mathrm{dom}(f)$ and $\mathbf{v}$ such that $\mathbf{x} + t\mathbf{v} \in \mathrm{dom}(f)$.

When $M = 2$, the function f is said to be standard self-concordant.

Definition (Self-concordant barrier)

A standard self-concordant function f is a ν-self-concordant barrier of the set $\mathrm{dom}(f)$ with parameter $\nu > 0$ if

$$\sup_{\mathbf{u} \in \mathbb{R}^p} \left\{ 2\mathbf{u}^T \nabla f(\mathbf{x}) - \mathbf{u}\nabla^2 f(\mathbf{x})\mathbf{u} \right\} \leq \nu, \quad \forall \mathbf{x} \in \mathrm{dom}(f).$$

Example

- $f(\mathbf{x}) := -\sum_{i=1}^p \ln(x_i)$ is a p-self-concordant barrier of $\mathbb{R}^p_{++}$.
- $f(\mathbf{X}) := -\ln\det(\mathbf{X})$ is a p-self-concordant barrier of $\mathbb{S}^p_{++}$.

Key estimates

Definition (Analytic center)

Let $b_{\mathcal{U}}$ be a self-concordant barrier of a convex set $\mathcal{U}$. The analytic center is defined as

$$\mathbf{u}_c := \arg\min_{\mathbf{u}\in\mathrm{int}(\mathcal{U})} b_{\mathcal{U}}(\mathbf{u}).$$

Convention: $b_{\mathcal{U}}(\mathbf{u}_c) = 0$; otherwise shift the original $b_{\mathcal{U}}$ by the constant $-b_{\mathcal{U}}(\mathbf{u}_c)$.

Theorem ([21])

Define $f_c(\mathbf{x}) = \mathbf{u}_c^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u})$*. For any* $\sigma > 0$*,* f_σ *is convex and*

$$f_\sigma(\mathbf{x}) \le f(\mathbf{x}) \le f_\sigma(\mathbf{x}) + \sigma\nu\left\{1 + \left[\ln\left(\frac{f(\mathbf{x}) - f_c(\mathbf{x})}{\sigma\nu}\right)\right]_+\right\},$$

where $[a]_+ := \max\{0, a\}$.

Observation: If $f(\mathbf{x}) - f_c(\mathbf{x}) \le \sigma\nu\exp(\rho)$, $|f(\mathbf{x}) - f_\sigma(\mathbf{x})| \le (1+\rho)\sigma\nu \to 0$ as $\sigma \downarrow 0$ with any $\rho \in \mathbb{R}$.

⋆ Differentiability

Theorem ([21])

The smoother f_σ *is differentiable in* $\mathrm{int}(\mathrm{dom}(f_\sigma))$ *and* $\nabla f_\sigma(\mathbf{x}) = \mathbf{A}^T\mathbf{u}^\star(\mathbf{x})$*.*

For any $\mathbf{x}, \mathbf{y} \in \mathrm{int}(\mathrm{dom}(f_\sigma))$*,*

$$\|\nabla f_\sigma(\mathbf{y}) - \nabla f_\sigma(\mathbf{x})\|_2 \leq \sigma^{-1} c_{\mathbf{A}}(\mathbf{y}) \left[c_{\mathbf{A}}(\mathbf{y}) + \|\nabla f_\sigma(\mathbf{y}) - \nabla f_\sigma(\mathbf{x})\|\right] \|\mathbf{y} - \mathbf{x}\|_2 \,,$$

where

$$c_{\mathbf{A}}(\mathbf{y}) := \left\|\mathbf{A}^T\nabla^2 b_{\mathcal{U}}(\mathbf{u}^\star(\mathbf{x}))\mathbf{A}\right\|_2^{1/2},$$
$$\mathbf{u}^\star(\mathbf{x}) := \arg\max_{\mathbf{u}\in\mathcal{U}} \left\{\mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \sigma b_{\mathcal{U}}(\mathbf{u})\right\}.$$

Observation: ∇f_σ is Lipschitz-like.

A gradient method for self-concordant barrier smoothing

Barrier smoothing with the gradient method

1. Give the smoothness parameter $\sigma > 0$ and an accuracy $\varepsilon > 0$. Choose $\mathbf{x}^0 \in \mathbb{R}^p$ as a starting point.

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

1. Calculate $\nabla f_\sigma(\mathbf{x}^k) := \mathbf{A}^T \mathbf{u}^\star(\mathbf{x}^k)$.
2. Compute $r_k := \left\|\nabla f_\sigma(\mathbf{x}^k)\right\|_2$ and $c^k_{\mathbf{A}} := c_{\mathbf{A}}(\mathbf{x}^k)$.
3. If $r_k \leq \varepsilon$, terminate.
4. Otherwise, update $\mathbf{x}^{k+1} := \mathbf{x}^k - \alpha_k \nabla f_\sigma(\mathbf{x}^k)$, where $\alpha_k := \sigma \left[c^k_{\mathbf{A}}\left(c^k_{\mathbf{A}} + r_k\right)\right]^{-1}$.

Observation: The step size α_k adapts to the local structure of f_σ.

Theorem (cf. [21] for details)

$$f_\sigma(\mathbf{x}^k) - f^\star_\sigma \leq \frac{4\overline{c_{\mathbf{A}}}^2 \left\|\mathbf{x}^0 - \mathbf{x}^\star_\sigma\right\|_2^2}{\sigma k},$$

where $\mathbf{x}^\star_\sigma := \arg\min_{\mathbf{x}} f(\mathbf{x})$, $f^\star_\sigma := f_\sigma(\mathbf{x}^\star_\sigma)$, *and* $\overline{c_{\mathbf{A}}}$ *is any upper bound of* $c_{\mathbf{A}}(\mathbf{x})$ *on* $\mathrm{dom}(f_\sigma)$.

Advantages of self-concordant barrier smoothing

Advantage 1: Faster convergence

The step size α_k adapts to the local structure of the smoother, and thus the algorithm can *converge fast*.

Recall: $\alpha_k \equiv 1/L_{f_\gamma}$ for Nesterov smoothing.

Advantage 2: Easier subproblems

The domain $\mathrm{dom}(b_{\mathcal{U}})$ is the interior of $\mathcal{U}$, meaning that solving for $\mathbf{u}^\star(\mathbf{x}^k)$ is equivalent to solving the *unconstrained optimization problem*

$$\mathbf{u}^\star(\mathbf{x}^k) := \arg\max_{\mathbf{u}} \left\{ \mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \sigma b_{\mathcal{U}}(\mathbf{u}) \right\}.$$

Recall: For Nesterov smoothing we have

$$\mathbf{u}^\star(\mathbf{x}^k) := \arg\max_{\mathbf{u}\in\mathcal{U}} \left\{ \mathbf{u}^T\mathbf{A}\mathbf{x} - f^*(\mathbf{u}) - \sigma p_{\mathcal{U}}(\mathbf{u}) \right\}.$$

Example: Quadratically constrained quadratic programming

Quadratically constrained quadratic programming (QCQP)

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{Q} \in \mathbb{R}^{m \times m}$ be positive semidefinite, $\mathbf{B} \in \mathbb{R}^{m \times m}$ be Hermitian positive definite, and $\mathbf{b} \in \mathbb{R}^m$. A QCQP problem takes the following form.

$$g^\star := \min_{\mathbf{y} \in \mathbb{R}^m} \left\{ \mathbf{y}^T \mathbf{Q} \mathbf{y} + \mathbf{b}^T \mathbf{y} : \mathbf{y}^T \mathbf{B} \mathbf{y} \leq 1, \mathbf{A}^T \mathbf{y} = 0 \right\}.$$

The equivalent dual form of QCQP is the following.

$$f^\star := \min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) := \max_{\mathbf{u}} \left\{ \mathbf{u}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) - \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} : \mathbf{u} \in \mathcal{U} \right\} \right\},$$

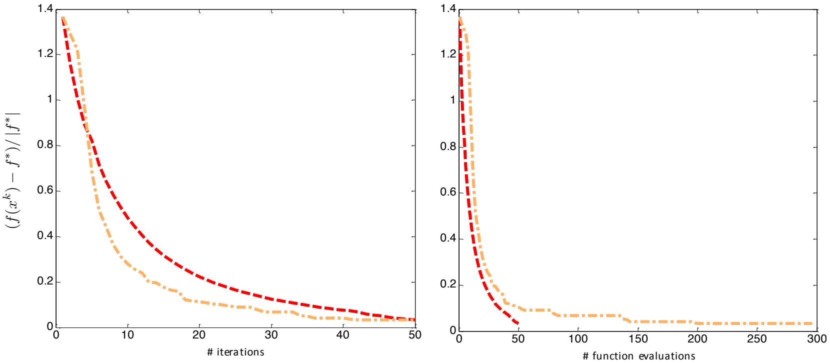
where $\mathcal{U} := \left\{ \mathbf{u} : \mathbf{u}^T \mathbf{B} \mathbf{u} \leq 1, \mathbf{u} \in \mathbb{R}^m \right\}$.

Observation: When $\mathbf{Q}$ is singular, f is nonsmooth.

Two approaches to solve the dual form of QCQP

1. **Nesterov smoothing:** Choose the prox-function $p_{\mathcal{U}}(\mathbf{u}) := \frac{1}{2} \mathbf{u}^T \mathbf{B} \mathbf{u}$.
2. **Barrier smoothing:** Choose the self-concordant barrier $b_{\mathcal{U}}(\mathbf{u}) := -\ln\left(1 - \mathbf{u}^T \mathbf{B} \mathbf{u}\right)$.

Numerical result



Orange: Nesterov smoothing *with line search*; Red: Barrier smoothing

References

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

[2] T. W. Anderson, I. Olkin, and L. G. Underhill.
Generation of random orthogonal matrices.
SIAM J. Sci. Stat. Comput., 8(4):625–629, 1987.

[3] Emmanuel Candès, Xiaodong Li, Yi Ma, and John Wright.
Robust principal component analysis?
J. ACM, 58(3), may 2011.

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

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

References

[6] David Donoho and Jared Tanner.
Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing.
Phil. Trans. R. Soc. A, 367:4273–4293, 2009.

[7] David L. Donoho.
Neighborly polytopes and sparse solution of underdetermined linear equations.
Technical report, Stanford University, 2004.

[8] David L. Donoho and Xiaoming Huo.
Uncertainty principles and ideal atomic decomposition.
IEEE Trans. Inf. Theory, 47(7):2845–2862, November 2001.

[9] David L. Donoho and Philip B. Stark.
Uncertainty principles and signal recovery.
SIAM J. Appl. Math., 49(3):906–931, June 1989.

[10] David L. Donoho and Jared Tanner.
Counting faces of randomly projected polytopes when the projection radically lowers dimension.
J. Amer. Math. Soc., 22(1):1–53, January 2009.

References

[11] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho.
Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA).
Appl. Comput. Harmon. Anal., 19:340–358, 2005.

[12] Michael Elad and Alfred M. Bruckstein.
A generalized uncertainty principle and sparse representation in pairs of bases.
IEEE Trans. Inf. Theory, 48(9):2558–2567, September 2002.

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

[14] Rina Foygel and Lester Mackey.
Corrupted sensing: Novel guarantees for separating structured signals.
2014.
arXiv:1305.2524v2 [cs.IT].

[15] Daniel A. Klain and Gian-Carlo Rota.
Introduction to Geometric Probability.
Cambridge Univ. Press, Cambridge, UK, 1997.

References

[16] Michael B. McCoy.
A geometric analysis of convex demixing.
PhD thesis, California Institute of Technology, 2013.

[17] Michael B. McCoy, Volkan Cevher, Quoc Tran-Dinh, Afsaneh Asaei, and Luca Baldassarre.
Convexity in source separation: Models, geometry, and algorithms.
IEEE Signal Process. Mag., 31(3):87–95, May 2014.

[18] Michael B. McCoy and Joel A. Tropp.
Sharp recovery bounds for convex demixing, with applications.
Found. Comput. Math., 14:503–567, 2014.

[19] Vitali D. Milman and Gideon Schechtman.
Asymptotic Theory of Finite Dimensional Normed Spaces.
Springer-Verl., second edition, 2001.

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

[21] Tran-Dinh Quoc, Yen-Huan Li, and Volkan Cevher.
Barrier smoothing for nonsmooth convex minimization.
In *2014 IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 1503–1507, 2014.

References

[22] Ghristoph Studer, Patrick Kuppinger, Graeme Pope, and Helmut Bölcskei.
Recovery of sparsely corrupted signals.
IEEE Trans. Inf. Theory, 58(5):3115–3130, May 2012.

[23] L. R. Welch.
Lower bounds on the maximum cross correlation of signals.
IEEE Trans. Inf. Theory, IT-20(3):397–399, May 1974.

[24] Hermann Weyl.
The Theory of Groups and Quantum Mechanics.
Dover, 1950.