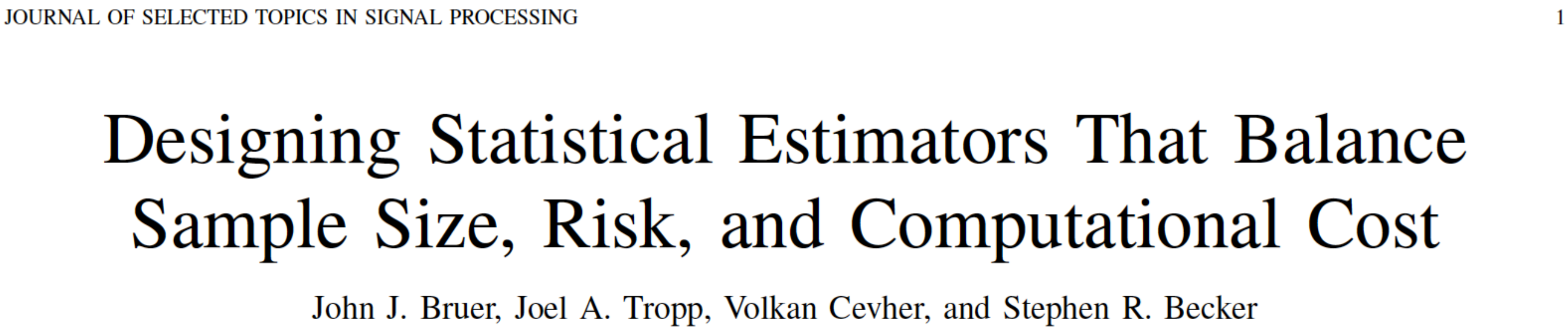
# Exploiting Large Sample Size to Reduce Statistical Risk and Computational Cost

Harlin Lee

March 26, 2018

18.898G Midterm Presentation



JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING 1

# Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost

John J. Bruer, Joel A. Tropp, Volkan Cevher, and Stephen R. Becker

[1] Paper from 2015
By people in Caltech, EPFL & UCBoulder

# Motivation: Data as a Computational Resource?

- Many problems nowadays involve massive data sets.
- Surprisingly, increase in data size doesn’t always lead to higher computational cost. SVM example in 2008 [2]
- This paper proposes an approach to:
  - systematically take advantage of large sample size in solving statistical problems through convex optimization.
  - by using extra samples for smoothing in the dual domain.

# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)

# Regularized Linear Regression Problem

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{\natural} + \boldsymbol{v},$$

$\boldsymbol{x}^{\natural} \in \mathbb{R}^{d}$, *d* parameters of a statistical model

$\boldsymbol{b} \in \mathbb{R}^{m}$ *m* observations

$\boldsymbol{A} \in \mathbb{R}^{m \times d}$ *m* *d*-dimensional inputs

$\boldsymbol{v} \in \mathbb{R}^{m}$ i.i.d. zero-mean noise

# Regularized Linear Regression Problem

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{\natural} + \boldsymbol{v}$$

| | |
|---|---|
| $\boldsymbol{x}^{\natural} \in \mathbb{R}^{d},$ | *d* parameters of a statistical model |
| $\boldsymbol{b} \in \mathbb{R}^{m}$ | *m* observations |
| $\boldsymbol{A} \in \mathbb{R}^{m \times d}$ | *m d*-dimensional inputs |
| $\boldsymbol{v} \in \mathbb{R}^{m}$ | i.i.d. zero-mean noise |

$$\widehat{\boldsymbol{x}} := \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \leq \sqrt{m \cdot R_{\max}} =: \epsilon$$

# Regularized Linear Regression Problem

$\boldsymbol{x}^{\natural} \in \mathbb{R}^d$, $d$ parameters of a statistical model

$\boldsymbol{b} \in \mathbb{R}^m$ $m$ observations

$\boldsymbol{A} \in \mathbb{R}^{m \times d}$ $m$ $d$-dimensional inputs

$\boldsymbol{v} \in \mathbb{R}^m$ i.i.d. zero-mean noise

$$\boldsymbol{b} = \boldsymbol{A}\boldsymbol{x}^{\natural} + \boldsymbol{v}$$

$f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ Convex regularizer function, i.e. L1 norm

$$\widehat{\boldsymbol{x}} := \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

$$\text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \leq \sqrt{m \cdot R_{\max}} =: \epsilon$$

$R_{\max}$ Maximal empirical risk that's tolerated

# Statistical and Empirical Risk

$$R(\widehat{\boldsymbol{x}}) = \frac{1}{m}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^{\natural}\|^2$$

Average squared prediction error for sample size *m*

$$\mathbb{E}_{\boldsymbol{v}}[R(\widehat{\boldsymbol{x}})]$$

Statistical risk of estimate $\widehat{\boldsymbol{x}}$

# Statistical and Empirical Risk

$$R(\widehat{\boldsymbol{x}}) = \frac{1}{m}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{A}\boldsymbol{x}^{\natural}\|^2$$

Average squared prediction error for sample size *m*

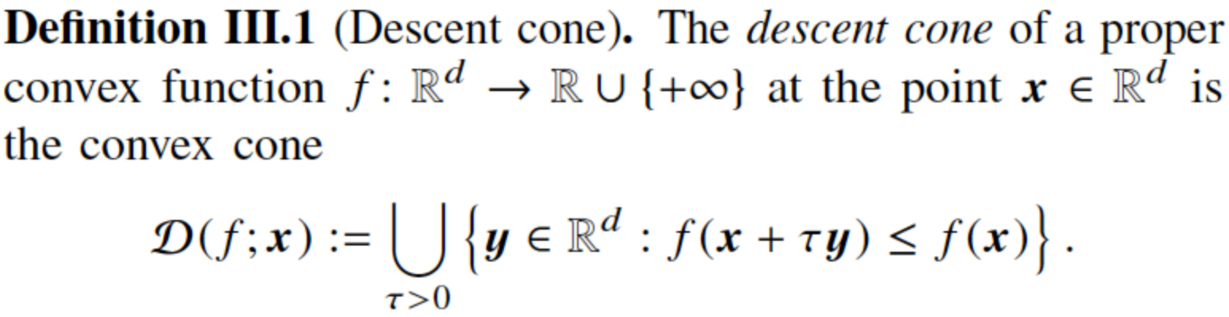
$$\mathbb{E}_{\boldsymbol{v}}[R(\widehat{\boldsymbol{x}})]$$

Statistical risk of estimate $\widehat{\boldsymbol{x}}$

$$\widehat{R}(\widehat{\boldsymbol{x}}) := \frac{1}{m}\|\boldsymbol{A}\widehat{\boldsymbol{x}} - \boldsymbol{b}\|^2$$

Empirical risk of estimate $\widehat{\boldsymbol{x}}$

# Statistical Dimension of Descent Cone

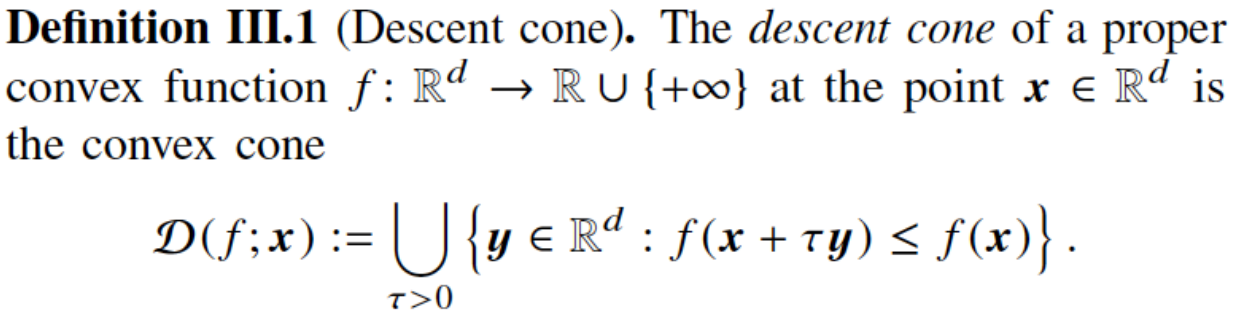


**Definition III.1** (Descent cone)**.** The *descent cone* of a proper convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ at the point $\boldsymbol{x} \in \mathbb{R}^d$ is the convex cone

$$\mathcal{D}(f; \boldsymbol{x}) := \bigcup_{\tau > 0} \left\{ \boldsymbol{y} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau \boldsymbol{y}) \leq f(\boldsymbol{x}) \right\}.$$

= set of directions that decrease $f$ locally at $\mathbf{x}$

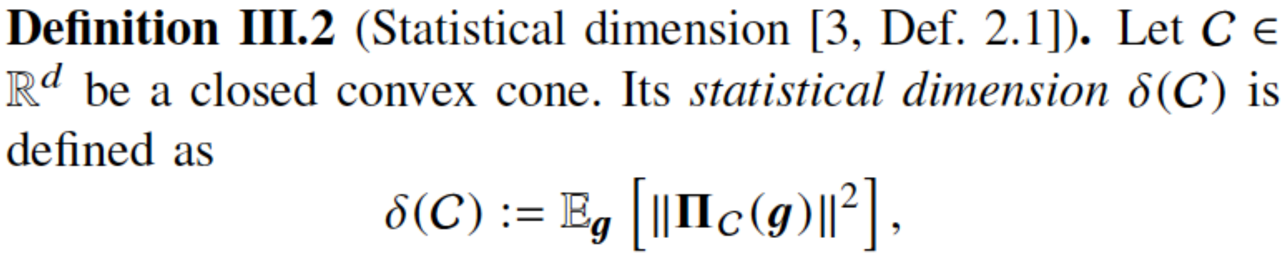
# Statistical Dimension of Descent Cone



**Definition III.1** (Descent cone)**.** The *descent cone* of a proper convex function $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ at the point $\boldsymbol{x} \in \mathbb{R}^d$ is the convex cone

$$\mathcal{D}(f;\boldsymbol{x}) := \bigcup_{\tau>0} \left\{ \boldsymbol{y} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau \boldsymbol{y}) \leq f(\boldsymbol{x}) \right\} .$$

= set of directions that decrease $f$ locally at $\mathbf{x}$



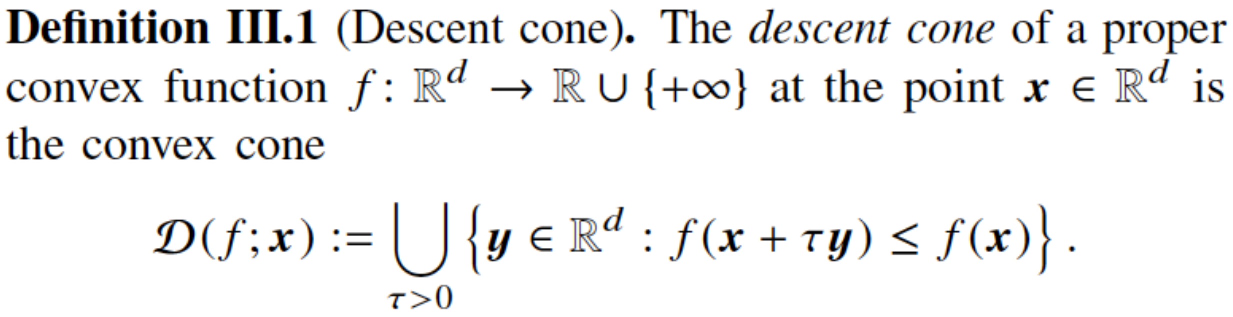
**Definition III.2** (Statistical dimension [3, Def. 2.1])**.** Let $C \in \mathbb{R}^d$ be a closed convex cone. Its *statistical dimension* $\delta(C)$ is defined as

$$\delta(C) := \mathbb{E}_{\boldsymbol{g}} \left[ \|\boldsymbol{\Pi}_C(\boldsymbol{g})\|^2 \right],$$

| | |
|---|---|
| $\boldsymbol{g} \in \mathbb{R}^d$ | i.i.d Gaussian noise |
| $\boldsymbol{\Pi}_C$ | Projection onto $C$ |

# Statistical Dimension of Descent Cone

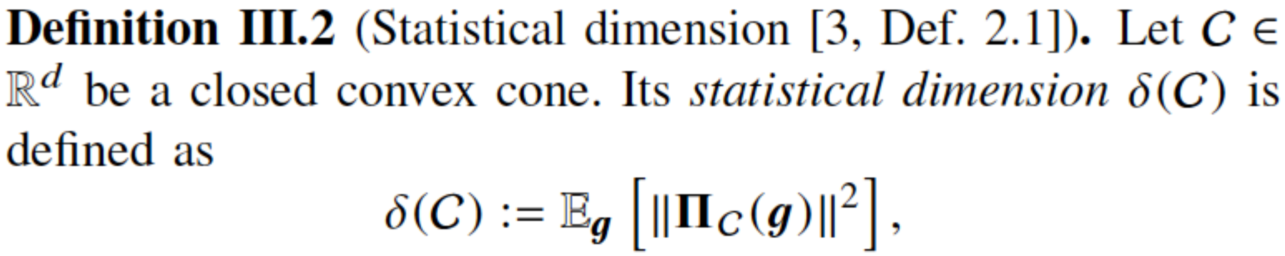
$$\boxed{\delta(\mathcal{D}(f;\boldsymbol{x}^{\natural}))}$$



**Definition III.1** (Descent cone)**.** The *descent cone* of a proper convex function $f\colon \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ at the point $\boldsymbol{x} \in \mathbb{R}^d$ is the convex cone

$$\mathcal{D}(f;\boldsymbol{x}) := \bigcup_{\tau>0} \left\{\boldsymbol{y} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau\boldsymbol{y}) \le f(\boldsymbol{x})\right\}.$$

= set of directions that decrease $f$ locally at **x**



**Definition III.2** (Statistical dimension [3, Def. 2.1])**.** Let $C \in \mathbb{R}^d$ be a closed convex cone. Its *statistical dimension* $\delta(C)$ is defined as

$$\delta(C) := \mathbb{E}_{\boldsymbol{g}}\left[\|\boldsymbol{\Pi}_C(\boldsymbol{g})\|^2\right],$$

| | |
|---|---|
| $\boldsymbol{g} \in \mathbb{R}^d$ | i.i.d Gaussian noise |
| $\boldsymbol{\Pi}_C$ | Projection onto $C$ |

# Roadmap

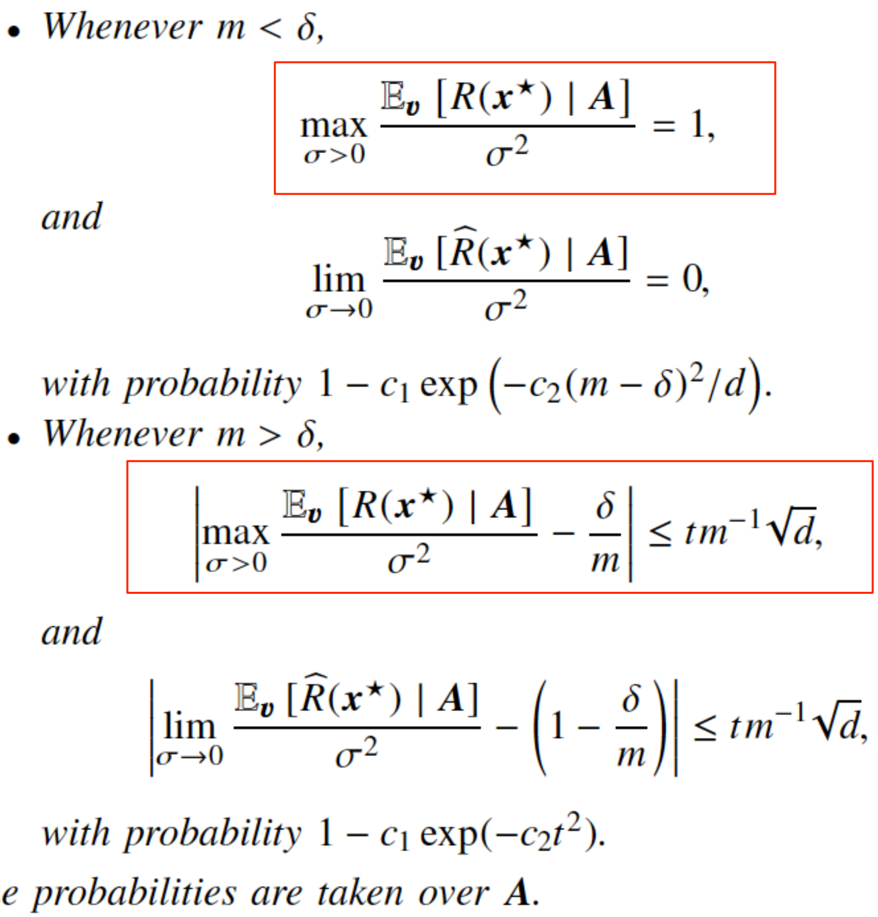
1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)
2. Show relationship between R, sample size (m), and δ

[3], [4]

If **A** has orthonormal rows,
$\sigma$ is standard deviation of $\boldsymbol{v}$,
**x*** is minimizer of our regression problem,
and $c_1$, $c_2$ are some constants:

[3], [4]

If **A** has orthonormal rows,
$\sigma$ is standard deviation of $\boldsymbol{v}$,
**x*** is minimizer of our regression problem,
and $c_1$, $c_2$ are some constants:



- *Whenever* $m < \delta$,

$$\boxed{\max_{\sigma>0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^{\star}) \mid \boldsymbol{A}\right]}{\sigma^2} = 1,}$$

*and*

$$\lim_{\sigma\to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^{\star}) \mid \boldsymbol{A}\right]}{\sigma^2} = 0,$$

*with probability* $1 - c_1 \exp\left(-c_2(m-\delta)^2/d\right)$.

- *Whenever* $m > \delta$,

$$\boxed{\left|\max_{\sigma>0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^{\star}) \mid \boldsymbol{A}\right]}{\sigma^2} - \frac{\delta}{m}\right| \le t m^{-1}\sqrt{d},}$$

*and*

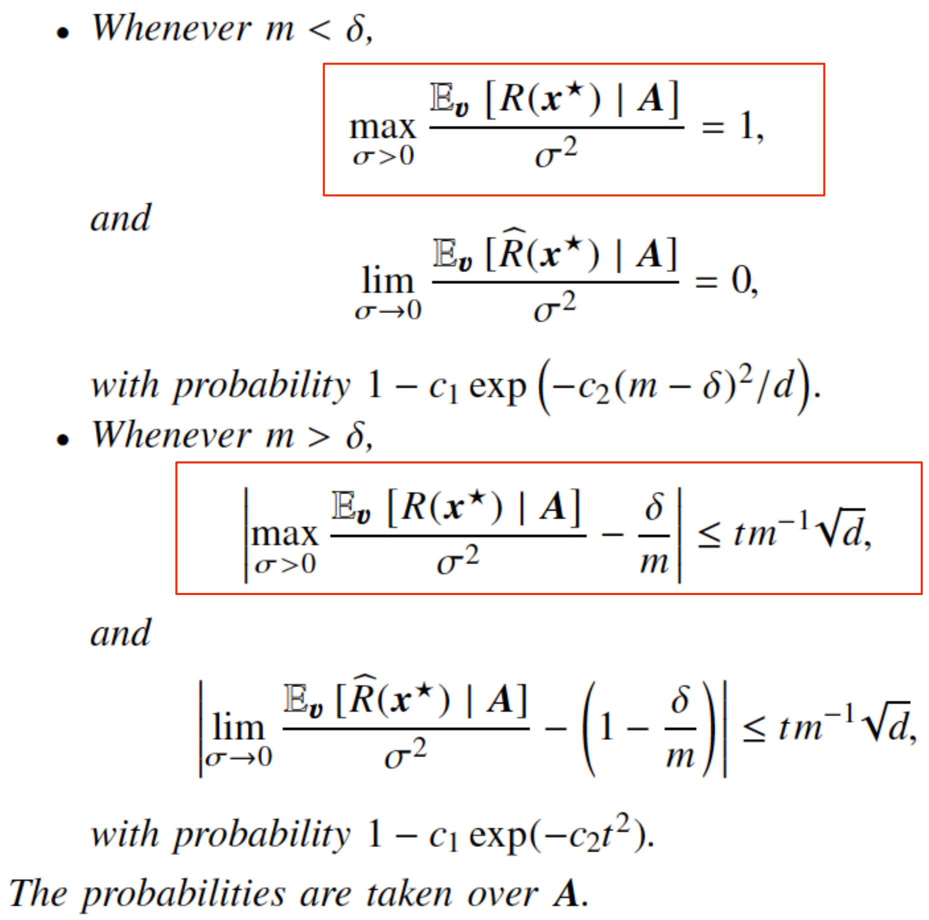
$$\left|\lim_{\sigma\to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^{\star}) \mid \boldsymbol{A}\right]}{\sigma^2} - \left(1 - \frac{\delta}{m}\right)\right| \le t m^{-1}\sqrt{d},$$

*with probability* $1 - c_1 \exp(-c_2 t^2)$.

*The probabilities are taken over* $\boldsymbol{A}$.

[3], [4]

If **A** has orthonormal rows,
$\sigma$ is standard deviation of $\boldsymbol{v}$,
**x*** is minimizer of our regression problem,
and $c_1$, $c_2$ are some constants:



- *Whenever* $m < \delta$,

$$\boxed{\max_{\sigma>0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} = 1,}$$

*and*

$$\lim_{\sigma\to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} = 0,$$

*with probability* $1 - c_1 \exp\left(-c_2(m-\delta)^2/d\right)$.

- *Whenever* $m > \delta$,

$$\boxed{\left|\max_{\sigma>0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} - \frac{\delta}{m}\right| \le tm^{-1}\sqrt{d},}$$

*and*

$$\left|\lim_{\sigma\to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} - \left(1 - \frac{\delta}{m}\right)\right| \le tm^{-1}\sqrt{d},$$

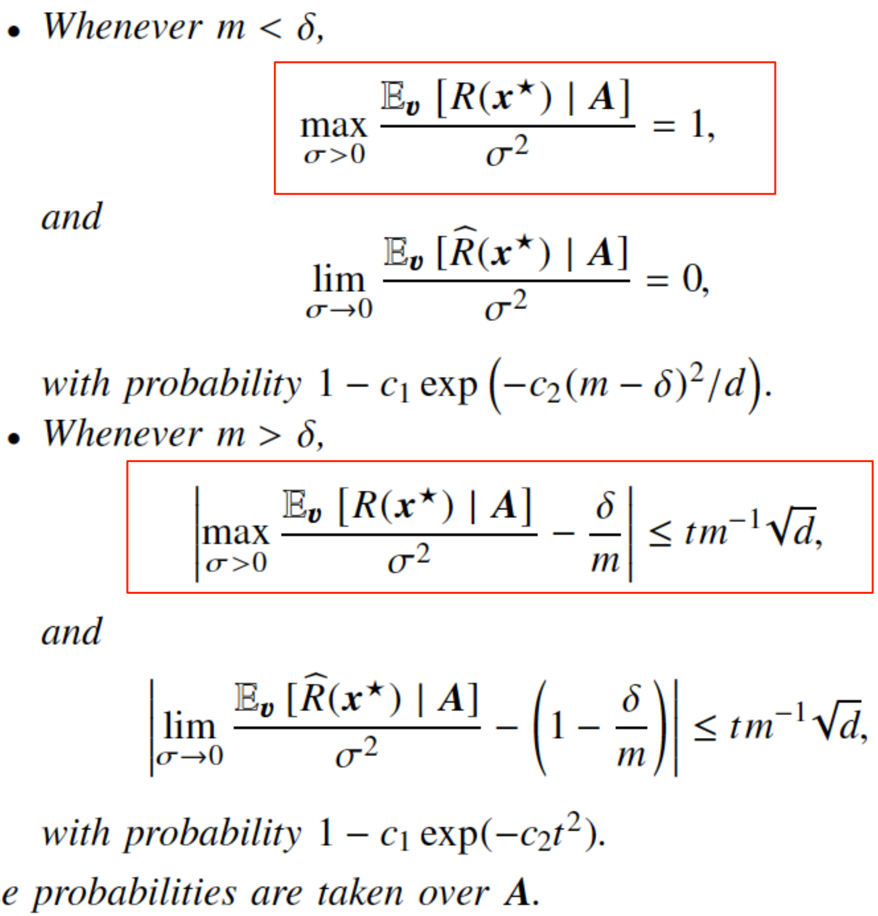
*with probability* $1 - c_1 \exp(-c_2 t^2)$.

*The probabilities are taken over* ***A***.

Worst-case statistical risk (R) can be bounded!

[3], [4]

If **A** has orthonormal rows,
$\sigma$ is standard deviation of ***v***,
**x*** is minimizer of our regression problem,
and $c_1$, $c_2$ are some constants:



- *Whenever* $m < \delta$,

$$\boxed{\max_{\sigma > 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} = 1,}$$

*and*

$$\lim_{\sigma \to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} = 0,$$

*with probability* $1 - c_1 \exp\left(-c_2(m-\delta)^2/d\right)$.

- *Whenever* $m > \delta$,

$$\boxed{\left|\max_{\sigma > 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[R(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} - \frac{\delta}{m}\right| \leq t m^{-1}\sqrt{d},}$$

*and*

$$\left|\lim_{\sigma \to 0} \frac{\mathbb{E}_{\boldsymbol{v}}\left[\widehat{R}(\boldsymbol{x}^\star) \mid \boldsymbol{A}\right]}{\sigma^2} - \left(1 - \frac{\delta}{m}\right)\right| \leq t m^{-1}\sqrt{d},$$

*with probability* $1 - c_1 \exp(-c_2 t^2)$.

*The probabilities are taken over* ***A***.

When $m < \delta$, R is constant -> statistical accuracy doesn't improve with $m$.

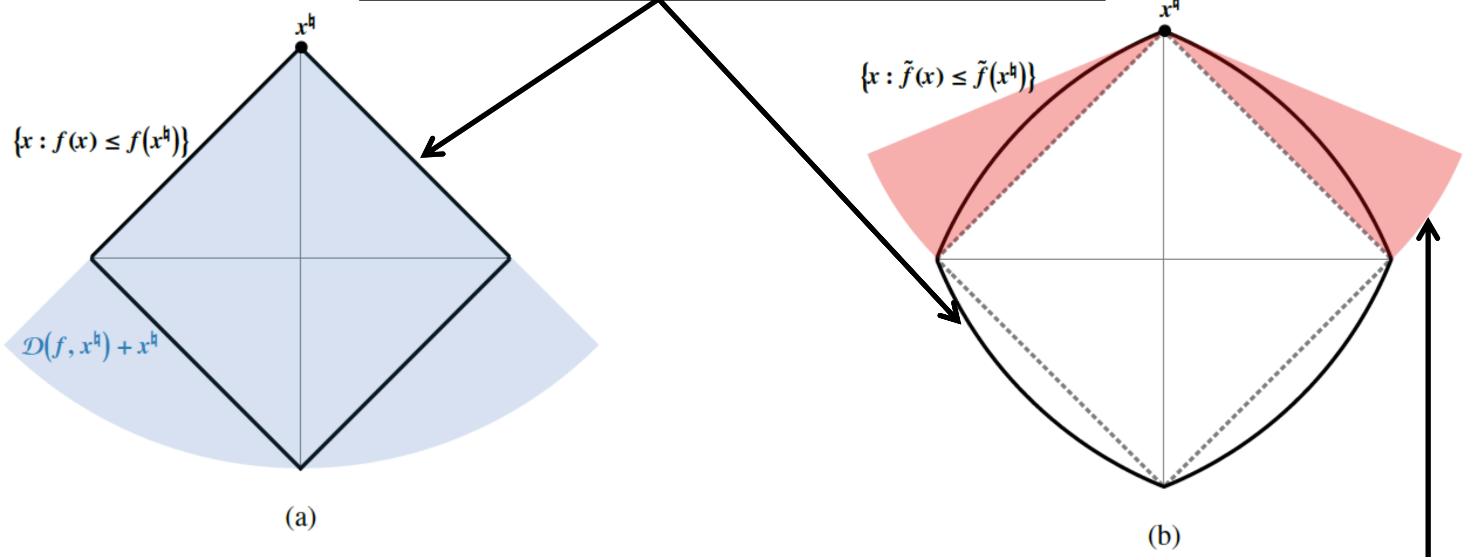
But after phase transition, worst-case risk decreases at $1/m$. That matches with our intuition.

# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone ($\delta$)
2. Show relationship between R, sample size (m), and $\delta$
3. Show geometric relationship between relaxing $f$, $\delta$, and R

# Geometry of relaxing *f* and increasing δ

Sublevel sets of *f* and relaxed *f*


$x^\natural$
$\{x : f(x) \leq f(x^\natural)\}$
$\mathcal{D}(f, x^\natural) + x^\natural$
$x^\natural$
$\{x : \tilde{f}(x) \leq \tilde{f}(x^\natural)\}$

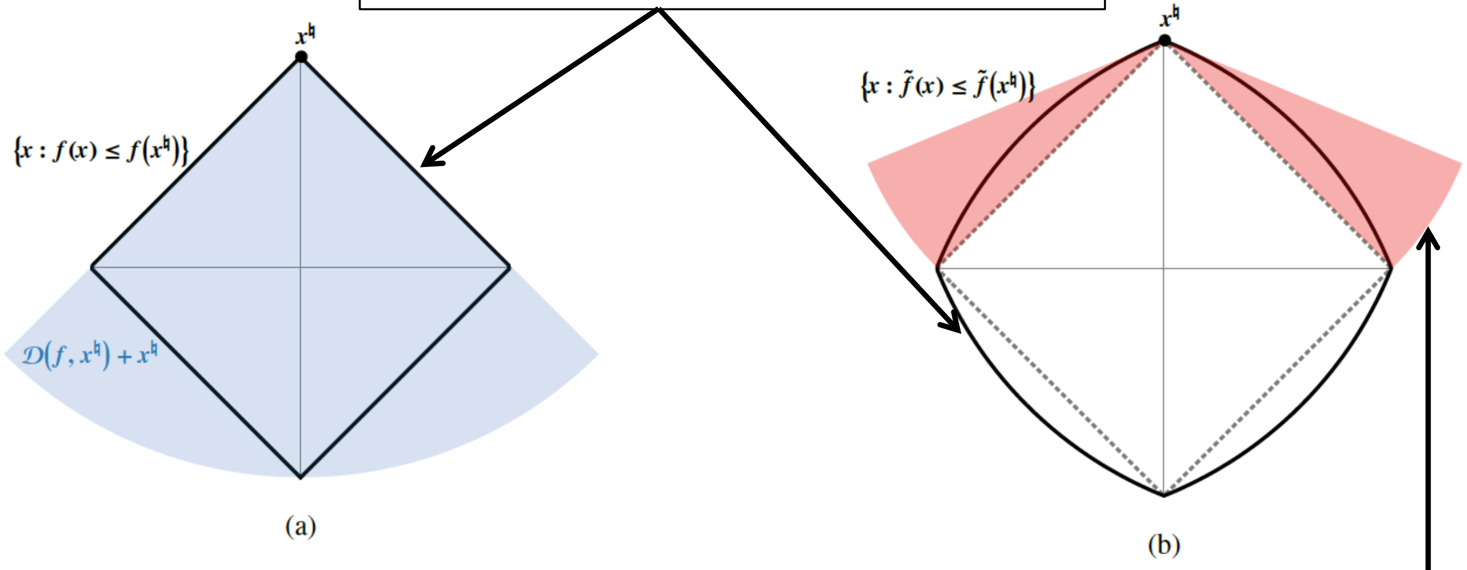
(a)

(b)

There is an increase of δ(D(relaxed f; ***x***♮))
-> increase in upper bound for R!

# Geometry of relaxing *f* and increasing δ

Sublevel sets of *f* and relaxed *f*



$x^\natural$
$\{x : f(x) \le f(x^\natural)\}$
$\mathcal{D}(f, x^\natural) + x^\natural$
(a)
$x^\natural$
$\{x : \tilde{f}(x) \le \tilde{f}(x^\natural)\}$
(b)

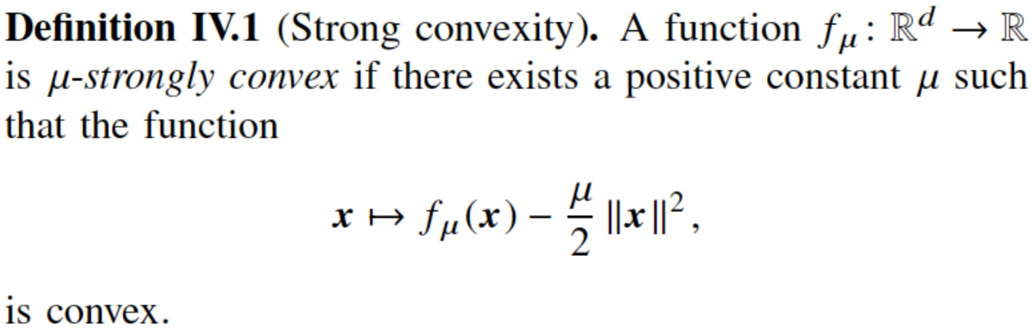
There is an increase of δ(D(relaxed f; ***x***♮))
-> increase in upper bound for R!

Relaxing *f* could help with optimization, BUT it decreases statistical accuracy. Note that if we smooth *f* directly, D(smoothed f; ***x***♮) becomes half-space and we lose control of δ. -> we start thinking about the dual problem.

# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)
2. Show relationship between risk, sample size (m), and δ
3. Show geometric relationship between relaxed *f*, δ, and R
4. Define strong convexity, Lipschitz gradient, and dual problem

# Strong Convexity and Lipschitz gradient



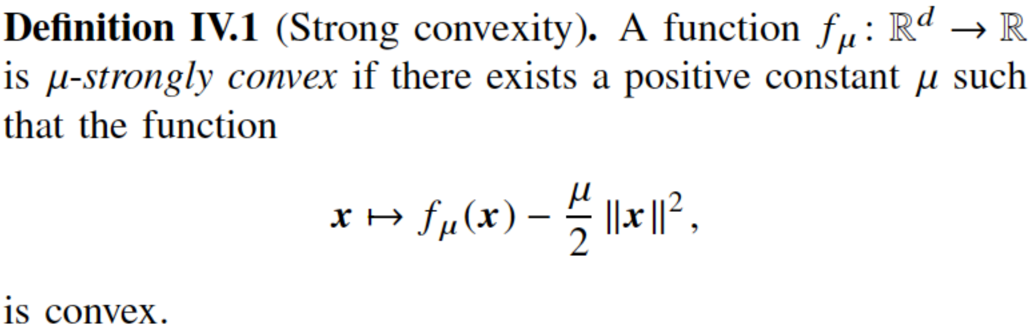
**Definition IV.1** (Strong convexity)**.** A function $f_\mu \colon \mathbb{R}^d \to \mathbb{R}$ is *$\mu$-strongly convex* if there exists a positive constant $\mu$ such that the function

$$\boldsymbol{x} \mapsto f_\mu(\boldsymbol{x}) - \frac{\mu}{2} \|\boldsymbol{x}\|^2,$$

is convex.

“More convex” functions have higher $\mu$

# Strong Convexity and Lipschitz gradient

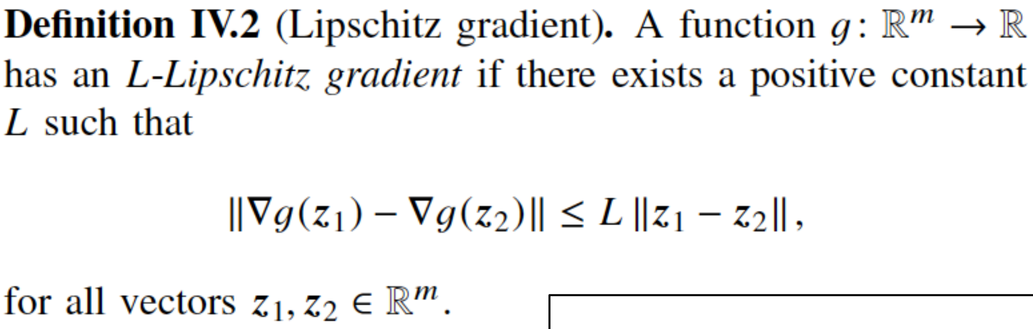


**Definition IV.1** (Strong convexity)**.** A function $f_\mu\colon \mathbb{R}^d \to \mathbb{R}$ is *$\mu$-strongly convex* if there exists a positive constant $\mu$ such that the function

$$\boldsymbol{x} \mapsto f_\mu(\boldsymbol{x}) - \frac{\mu}{2}\left\|\boldsymbol{x}\right\|^2,$$

is convex.

"More convex" functions have higher $\mu$



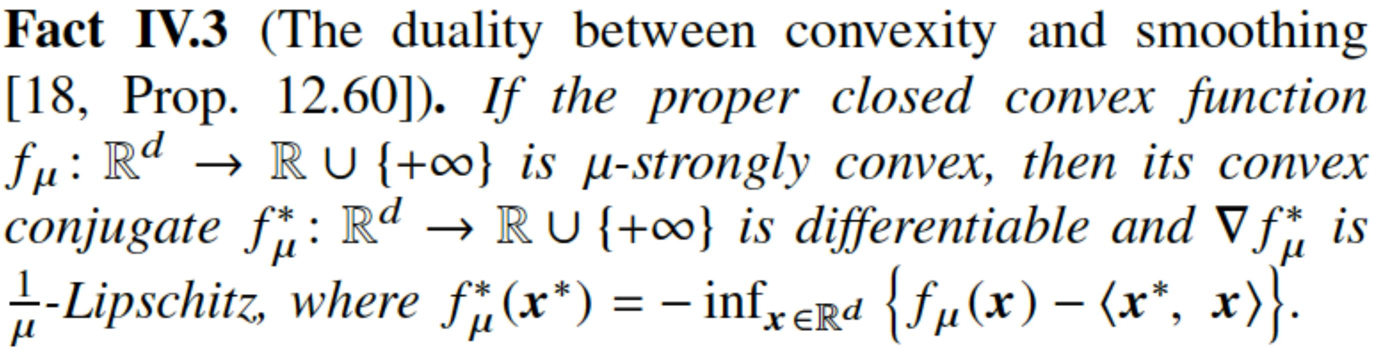
**Definition IV.2** (Lipschitz gradient)**.** A function $g\colon \mathbb{R}^m \to \mathbb{R}$ has an *$L$-Lipschitz gradient* if there exists a positive constant $L$ such that

$$\left\|\nabla g(z_1) - \nabla g(z_2)\right\| \leq L\left\|z_1 - z_2\right\|,$$

for all vectors $z_1, z_2 \in \mathbb{R}^m$.

Smoother functions have lower $L$

# Duality of Convexity and Smoothing



**Fact IV.3** (The duality between convexity and smoothing [18, Prop. 12.60]). *If the proper closed convex function* $f_\mu : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *is* $\mu$*-strongly convex, then its convex conjugate* $f_\mu^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ *is differentiable and* $\nabla f_\mu^*$ *is* $\frac{1}{\mu}$*-Lipschitz, where* $f_\mu^*(\boldsymbol{x}^*) = -\inf_{\boldsymbol{x} \in \mathbb{R}^d} \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{x}^*, \ \boldsymbol{x} \rangle \right\}$.

Higher $\mu$ -> more convex $f_\mu$ -> smoother conjugate $f_\mu{}^*$

Convex conjugate $f_\mu{}^*$ is also differentiable.

# Dual Problem of Our Regression

Primal: $$\widehat{\boldsymbol{x}}_\mu := \arg\min_{\boldsymbol{x}} f_\mu(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \leq \epsilon.$$

# Dual Problem of Our Regression

Primal: $$\widehat{\boldsymbol{x}}_\mu := \arg\min_{\boldsymbol{x}} f_\mu(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \le \epsilon.$$

Dual: $$\text{maximize} \quad g_\mu(z) := \inf_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle z,\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \rangle - \epsilon \|z\| \right\}$$

-> $$\begin{aligned} g_\mu(z) &= \inf_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{A}^T z,\ \boldsymbol{x} \rangle \right\} + \langle z,\ \boldsymbol{b} \rangle - \epsilon \|z\| \\ &= \underbrace{-f_\mu^*(\boldsymbol{A}^T z) + \langle z,\ \boldsymbol{b} \rangle}_{\text{smooth } \tilde{g}_\mu(z)} \underbrace{- \epsilon \|z\|}_{h(z) \text{ non-smooth}}, \end{aligned}$$

# Dual Problem of Our Regression

Primal:

$$\widehat{\boldsymbol{x}}_\mu := \arg\min_{\boldsymbol{x}} f_\mu(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \le \epsilon.$$

Dual:

$$\text{maximize} \quad g_\mu(z) := \inf_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle z,\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\rangle - \epsilon\,\|z\| \right\}$$

->

$$\begin{aligned} g_\mu(z) &= \inf_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{A}^T z,\ \boldsymbol{x}\rangle \right\} + \langle z,\ \boldsymbol{b}\rangle - \epsilon\,\|z\| \\ &= \underbrace{-f_\mu^*(\boldsymbol{A}^T z) + \langle z,\ \boldsymbol{b}\rangle}_{\text{smooth } \tilde{g}_\mu(z)} \underbrace{-\ \epsilon\,\|z\|}_{h(z) \text{ non-smooth}}, \end{aligned}$$

Gradient of dual problem:

$$\nabla \tilde{g}_\mu(z) = \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_z,$$

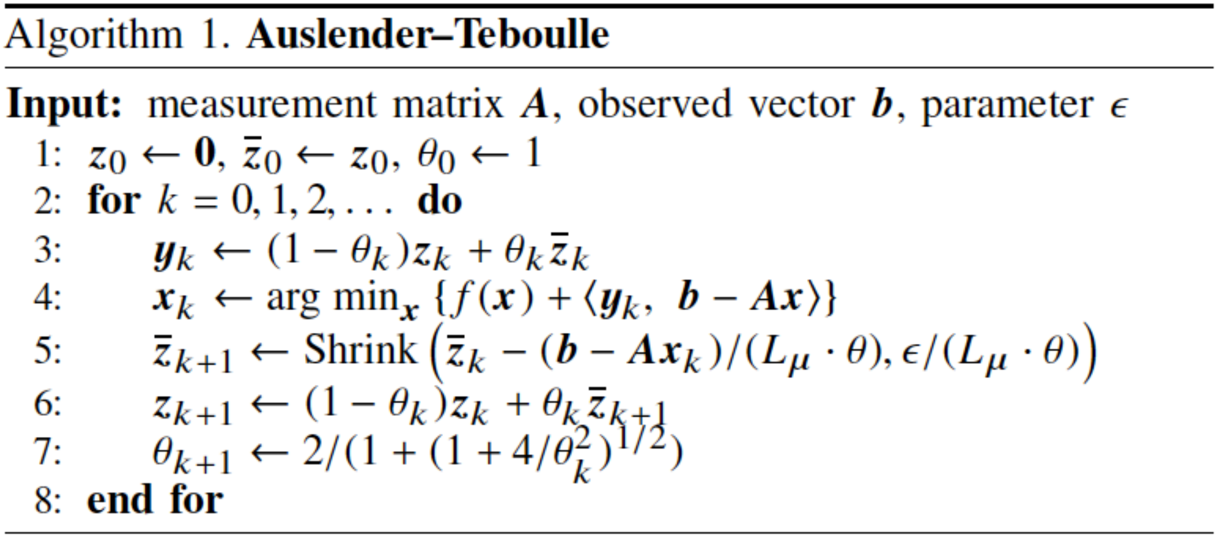
$$\boldsymbol{x}_z := \arg\min_{\boldsymbol{x}} \left\{ f_\mu(\boldsymbol{x}) - \langle \boldsymbol{A}^T z,\ \boldsymbol{x}\rangle \right\}$$

Subgradient of h(**z**) = $\varepsilon\mathbf{z}/\|\mathbf{z}\|$ if $\mathbf{z} \neq 0$, $\{\mathbf{y} \mid \|\mathbf{y}\| \le 1\}$ if $\mathbf{z} = 0$

# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)
2. Show relationship between risk, sample size (m), and δ
3. Show geometric relationship between relaxed *f*, δ, and R
4. Define strong convexity, Lipschitz gradient, and dual problem
5. Show relationship between dual-smoothing, computational cost, and risk

# Iterative First-Order Algorithm

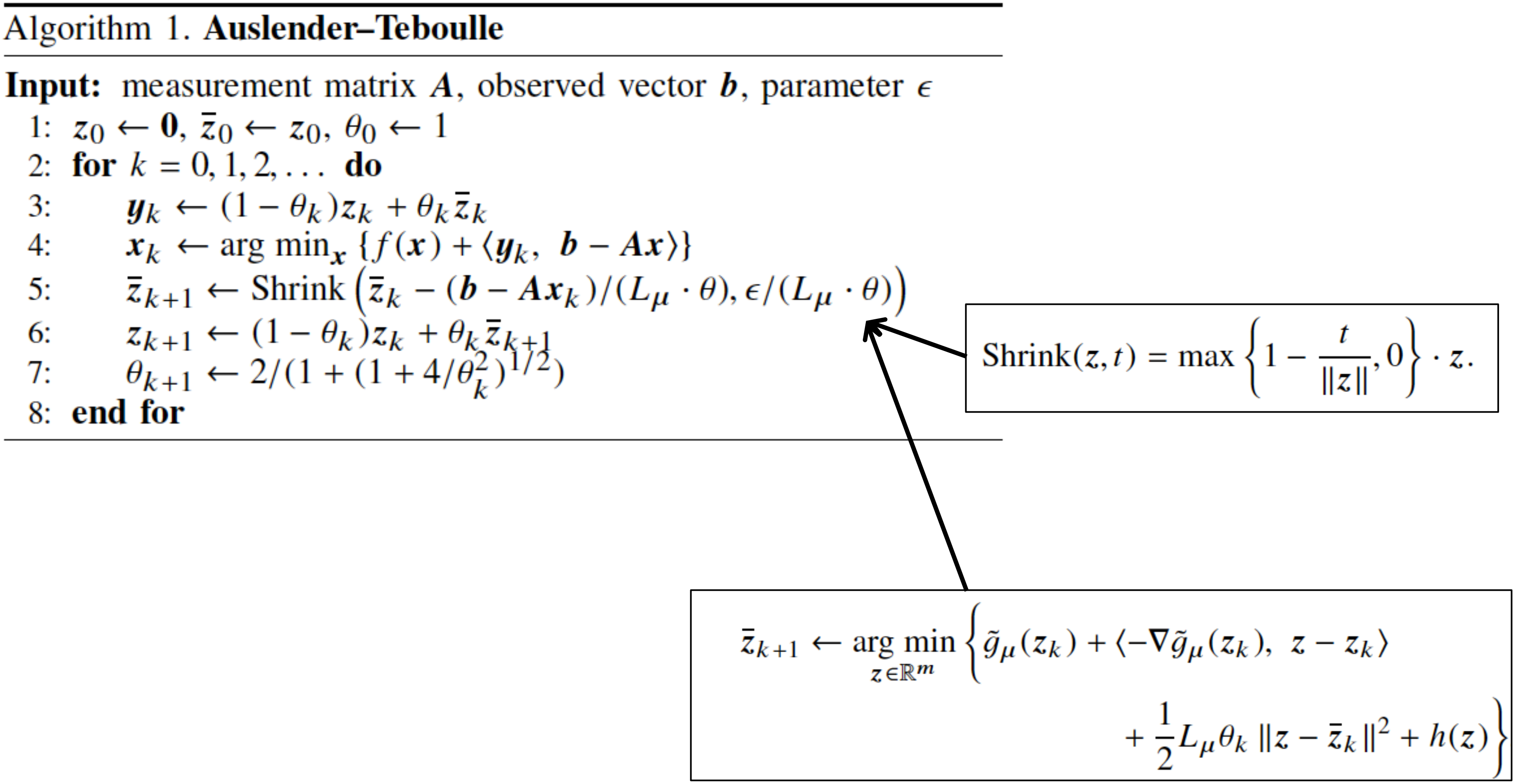


Algorithm 1. **Auslender–Teboulle**

```
Input: measurement matrix $A$, observed vector $b$, parameter $\epsilon$
$z_0 \leftarrow \mathbf{0}$, $\bar{z}_0 \leftarrow z_0$, $\theta_0 \leftarrow 1$
for $k = 0, 1, 2, \ldots$ do
    $\boldsymbol{y}_k \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_k$
    $\boldsymbol{x}_k \leftarrow \arg\min_{\boldsymbol{x}} \{ f(\boldsymbol{x}) + \langle \boldsymbol{y}_k, \ \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \rangle \}$
    $\bar{z}_{k+1} \leftarrow \text{Shrink}\left( \bar{z}_k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta) \right)$
    $z_{k+1} \leftarrow (1 - \theta_k) z_k + \theta_k \bar{z}_{k+1}$
    $\theta_{k+1} \leftarrow 2/(1 + (1 + 4/\theta_k^2)^{1/2})$
end for
```

[5], [6]

# Iterative First-Order Algorithm



Algorithm 1. **Auslender–Teboulle**

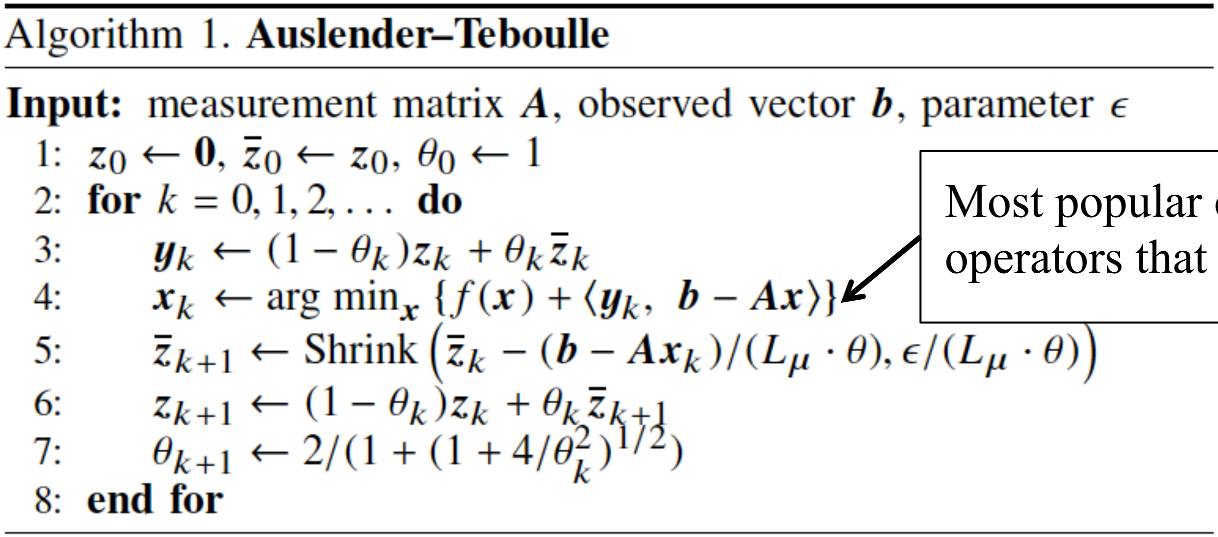
```
**Input:** measurement matrix $\boldsymbol{A}$, observed vector $\boldsymbol{b}$, parameter $\epsilon$
$z_0 \leftarrow \mathbf{0},\ \bar{z}_0 \leftarrow z_0,\ \theta_0 \leftarrow 1$
**for** $k = 0, 1, 2, \ldots$ **do**
    $\boldsymbol{y}_k \leftarrow (1-\theta_k) z_k + \theta_k \bar{z}_k$
    $\boldsymbol{x}_k \leftarrow \arg\min_{\boldsymbol{x}} \{ f(\boldsymbol{x}) + \langle \boldsymbol{y}_k,\ \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \rangle \}$
    $\bar{z}_{k+1} \leftarrow \text{Shrink}\left( \bar{z}_k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta) \right)$
    $z_{k+1} \leftarrow (1-\theta_k) z_k + \theta_k \bar{z}_{k+1}$
    $\theta_{k+1} \leftarrow 2/(1 + (1 + 4/\theta_k^2)^{1/2})$
**end for**
```

$$\text{Shrink}(z, t) = \max\left\{ 1 - \frac{t}{\|z\|}, 0 \right\} \cdot z.$$

$$\bar{z}_{k+1} \leftarrow \arg\min_{z \in \mathbb{R}^m} \left\{ \tilde{g}_\mu(z_k) + \langle -\nabla \tilde{g}_\mu(z_k),\ z - z_k \rangle + \frac{1}{2} L_\mu \theta_k \, \|z - \bar{z}_k\|^2 + h(z) \right\}$$

[5], [6]

# Iterative First-Order Algorithm



Algorithm 1. **Auslender–Teboulle**

```
Input: measurement matrix $A$, observed vector $b$, parameter $\epsilon$
1: $z_0 \leftarrow \mathbf{0}$, $\bar{z}_0 \leftarrow z_0$, $\theta_0 \leftarrow 1$
2: for $k = 0, 1, 2, \ldots$ do
3:     $\boldsymbol{y}_k \leftarrow (1-\theta_k) z_k + \theta_k \bar{z}_k$
4:     $\boldsymbol{x}_k \leftarrow \arg\min_{\boldsymbol{x}} \{ f(\boldsymbol{x}) + \langle \boldsymbol{y}_k, \ \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x} \rangle \}$
5:     $\bar{z}_{k+1} \leftarrow \text{Shrink}\left( \bar{z}_k - (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k)/(L_\mu \cdot \theta), \epsilon/(L_\mu \cdot \theta) \right)$
6:     $z_{k+1} \leftarrow (1-\theta_k) z_k + \theta_k \bar{z}_{k+1}$
7:     $\theta_{k+1} \leftarrow 2/(1 + (1 + 4/\theta_k^2)^{1/2})$
8: end for
```

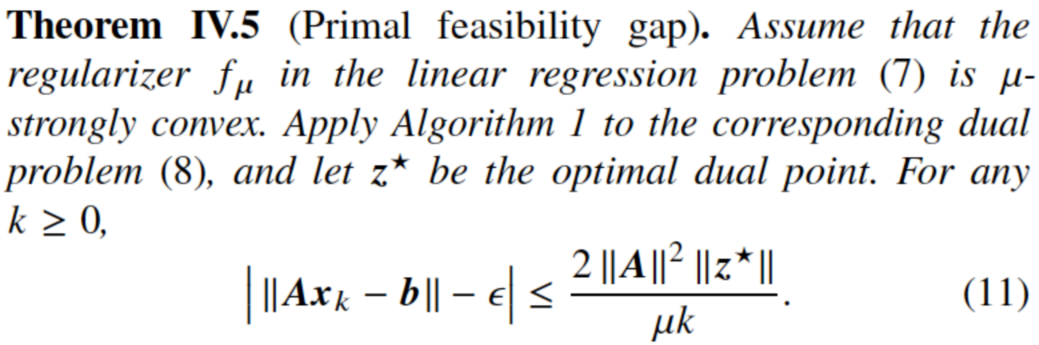
Most popular choices for *f* have proximal operators that are inexpensive to calculate

If line 4 costs little, then per iteration, cost will be O(*md*), for matrix multiplications with ***A***

[5], [6]

# Iterative First-Order Algorithm

Assuming dual-smoothing method + iterative first-order algorithm,



**Theorem IV.5** (Primal feasibility gap)**.** *Assume that the regularizer $f_\mu$ in the linear regression problem* (7) *is $\mu$-strongly convex. Apply Algorithm 1 to the corresponding dual problem* (8), *and let $\boldsymbol{z}^\star$ be the optimal dual point. For any* $k \geq 0$,

$$\left| \|\boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}\| - \epsilon \right| \leq \frac{2\,\|\boldsymbol{A}\|^2\,\|\boldsymbol{z}^\star\|}{\mu k}. \qquad (11)$$

If $\mu$ is large, upper bound for duality gap is lower.
Or for fixed ||Ax-b||=√($mR$) and ε, if $\mu$ is large*,* iteration $k$ can be smaller.

[7]

# Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost

Designing $\overset{=\widehat{x}}{\boxed{\text{Statistical Estimators}}}$ That Balance $\underset{=m}{\boxed{\text{Sample Size}}}$, $\underset{=R}{\boxed{\text{Risk}}}$, and $\underset{=O(kmd)}{\boxed{\text{Computational Cost}}}$

…using μ, dual-smoothing parameter

# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)
2. Show relationship between risk, sample size (m), and δ
3. Show geometric relationship between relaxed *f*, δ, and R
4. Define strong convexity, Lipschitz gradient, and dual problem
5. Show relationship between dual-smoothing, computational cost, and risk
6. Describe dual-smoothing approach

# How to relax regularizer $f$

Turn convex $f$ into strongly convex $f_\mu$ $$f_\mu(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|^2 .$$

# How to relax regularizer *f*

Turn convex *f* into strongly convex $f_\mu$ $\quad f_\mu(\boldsymbol{x}) := f(\boldsymbol{x}) + \frac{\mu}{2}\|\boldsymbol{x}\|^2 .$

Upper bound on number of iterations for convergence:

$$k \leq \frac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m} - \delta}.$$

If $\mu$ is higher, δ increases too, so there's probably an upper bound to how much we can increase $\mu$ before *k* stops decreasing.

If $\mu$ is constant, computational cost O(*kmd*) is proportional to √*m*.

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$ (phase transition theorem)

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$ (phase transition theorem)
- Computational cost is proportional to O(*kmd*) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}.$

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$ (phase transition)
- Computational cost is proportional to O($kmd$) and

$$k \le \frac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}.$$

$$\frac{\delta\left(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural)\right)}{m} = \frac{\bar{\delta}}{\bar{m} + (m-\bar{m})^\alpha}$$

α = 1:

α = -∞:

α in (-∞,1):

# Choosing *μ*, the Smoothing Parameter

Summary so far:

- *δ* is proportional to *μ* (geometric intuition)
- Worst case R is proportional to *δ*/*m*
- Computational cost is proportional to O(*kmd*) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}.$

Baseline δ from small baseline value for *μ*

$$\frac{\delta\left(\mathcal{D}(f_\mu; \boldsymbol{x}^\natural)\right)}{m} = \frac{\bar{\delta}}{\overline{m} + (m - \overline{m})^\alpha}$$

Baseline sample size *m*

α = 1:

α = -∞:

α in (-∞,1):

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$
- Computational cost is proportional to O($kmd$) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}.$

Baseline δ from small baseline value for $\mu$

$$? \; \frac{\delta\left(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural)\right)}{\uparrow\; m} = \frac{\bar{\delta}}{\overline{m} + (m-\overline{m})^\alpha}$$

Baseline sample size $m$

α = 1:

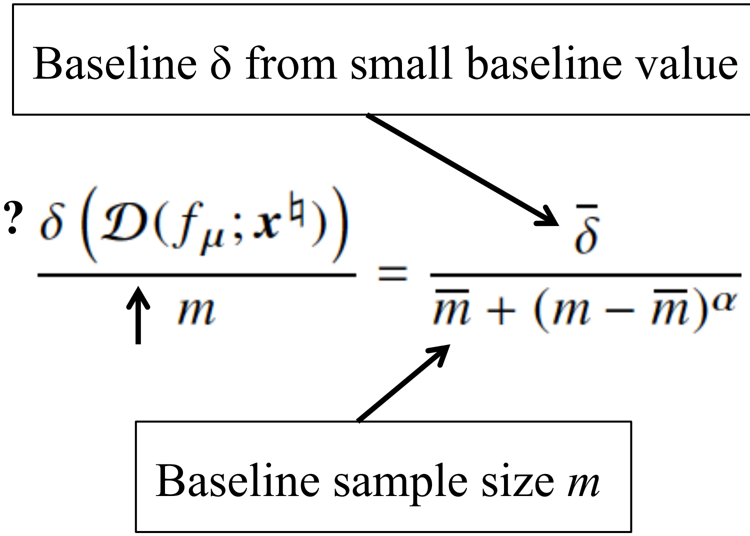
α = -∞:

α in (-∞,1):

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$
- Computational cost is proportional to O($kmd$) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}.$



Baseline δ from small baseline value for $\mu$

$$? \; \frac{\delta\left(\mathcal{D}(f_\mu; \boldsymbol{x}^\natural)\right)}{\uparrow \; m} = \frac{\bar{\delta}}{\overline{m} + (m - \overline{m})^\alpha}$$

Baseline sample size $m$

α = 1: $\mu$ is constant; δ is constant; R falls; cost increases

α = -∞:

α in (-∞,1):

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$
- Computational cost is proportional to O($kmd$) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m} - \delta}$.

Baseline δ from small baseline value for $\mu$

$$? \; \frac{\delta\left(\mathcal{D}(f_\mu;\boldsymbol{x}^\natural)\right)}{\uparrow \; m} = \frac{\bar{\delta}}{\overline{m} + (m - \overline{m})^\alpha}$$

Baseline sample size $m$

α = 1: $\mu$ is constant; δ is constant; R falls; cost increases  
α = -∞: $\mu$ increases; δ increases; R is constant; cost decreases  
α in (-∞,1):

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$
- Computational cost is proportional to O($kmd$) and $k \le \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}$.

Baseline δ from small baseline value for $\mu$

$$? \; \frac{\delta\left(\mathcal{D}(f_\mu; \boldsymbol{x}^\natural)\right)}{\uparrow\; m} = \frac{\bar{\delta}}{\overline{m} + (m - \overline{m})^{\alpha}}$$

Baseline sample size $m$

α = 1: $\mu$ is constant; δ is constant; R falls; cost increases

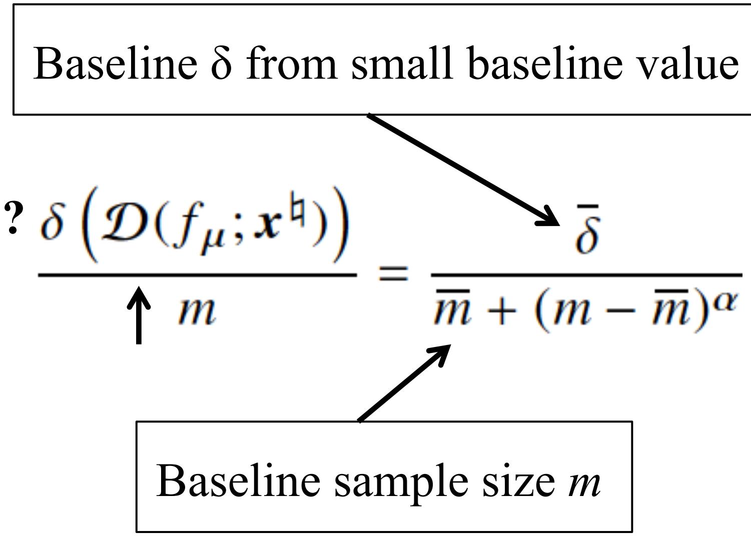
α = -∞: $\mu$ increases; δ increases; R is constant; cost decreases

α in (-∞,1): *balance!*

# Choosing $\mu$, the Smoothing Parameter

Summary so far:

- $\delta$ is proportional to $\mu$ (geometric intuition)
- Worst case R is proportional to $\delta/m$
- Computational cost is proportional to O($kmd$) and $k \leq \dfrac{2\,\|z^\star\|}{\gamma\mu\sigma\sqrt{m-\delta}}$.



Baseline δ from small baseline value for $\mu$

$$? \; \frac{\delta\left(\mathcal{D}(f_\mu; \boldsymbol{x}^\natural)\right)}{\uparrow\; m} = \frac{\bar{\delta}}{\overline{m} + (m - \overline{m})^{\alpha}}$$

Baseline sample size $m$

α = 1: $\mu$ is constant; δ is constant; R falls; cost increases

α = -∞: $\mu$ increases; δ increases; R is constant; cost decreases

α in (-∞,1): *balance!*

"When we have excess samples in the data set, we can exploit them to decrease the statistical risk of our estimator or to lower the computational cost through additional smoothing. A tradeoff arises from the balance between these two competing interests."

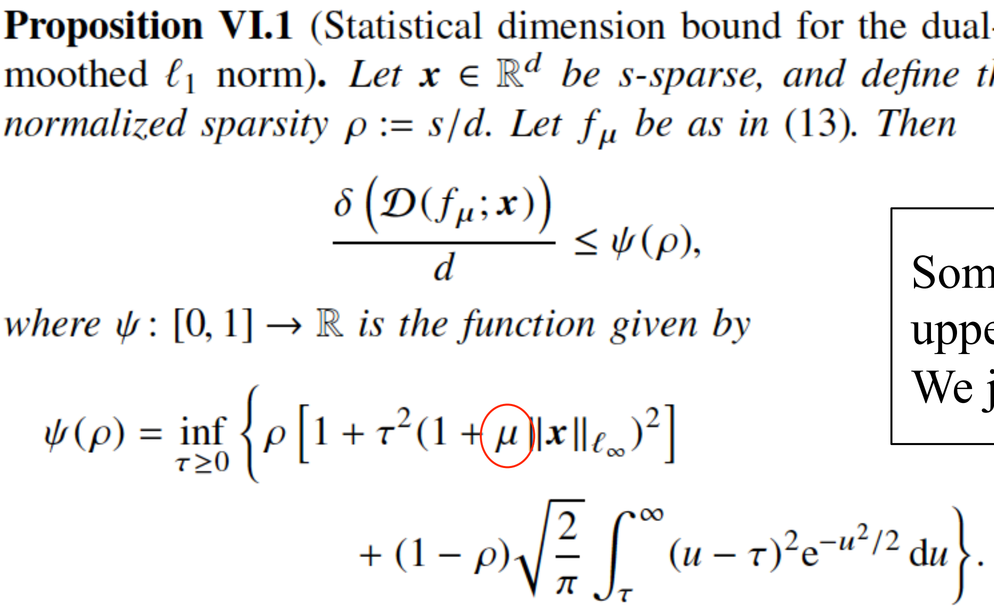
# Roadmap

1. Define the regression problem, risk (R), and statistical dimension of descent cone (δ)
2. Show relationship between risk, sample size (m), and δ
3. Show geometric relationship between relaxed *f*, δ, and R
4. Define strong convexity, Lipschitz gradient, and dual problem
5. Show relationship between smoothing, computational cost, and risk
6. Describe smoothing approach
7. Results on simulations

# LASSO

$$\widehat{\boldsymbol{x}}_\mu := \arg\min_{\boldsymbol{x}} f_\mu(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \leq \epsilon.$$

$$f_\mu(\boldsymbol{x}) = \|\boldsymbol{x}\|_{\ell_1} + \frac{\mu}{2}\|\boldsymbol{x}\|^2$$



**Proposition VI.1** (Statistical dimension bound for the dual-smoothed $\ell_1$ norm)**.** *Let* $\boldsymbol{x} \in \mathbb{R}^d$ *be s-sparse, and define the normalized sparsity* $\rho := s/d$. *Let* $f_\mu$ *be as in* (13). *Then*

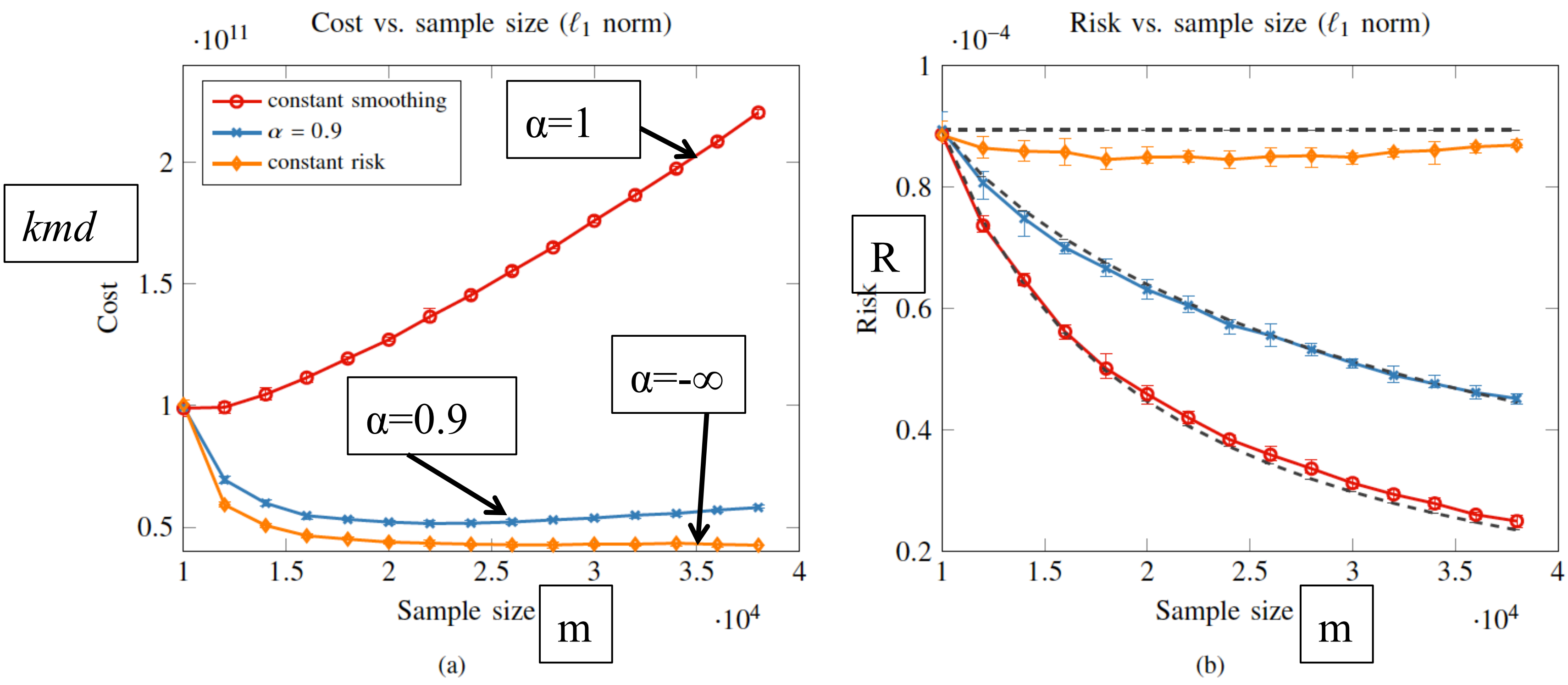
$$\frac{\delta\left(\mathcal{D}(f_\mu; \boldsymbol{x})\right)}{d} \leq \psi(\rho),$$

*where* $\psi : [0, 1] \to \mathbb{R}$ *is the function given by*

$$\psi(\rho) = \inf_{\tau \geq 0}\left\{\rho\left[1 + \tau^2(1 + \mu\|\boldsymbol{x}\|_{\ell_\infty})^2\right] + (1 - \rho)\sqrt{\frac{2}{\pi}}\int_\tau^\infty (u - \tau)^2 \mathrm{e}^{-u^2/2}\,\mathrm{d}u\right\}.$$

Some complicated looking value for upper bound of δ.
We just need to notice $\mu$ there.

# LASSO



*kmd*

α=1

α=0.9

α=-∞

m

(a)

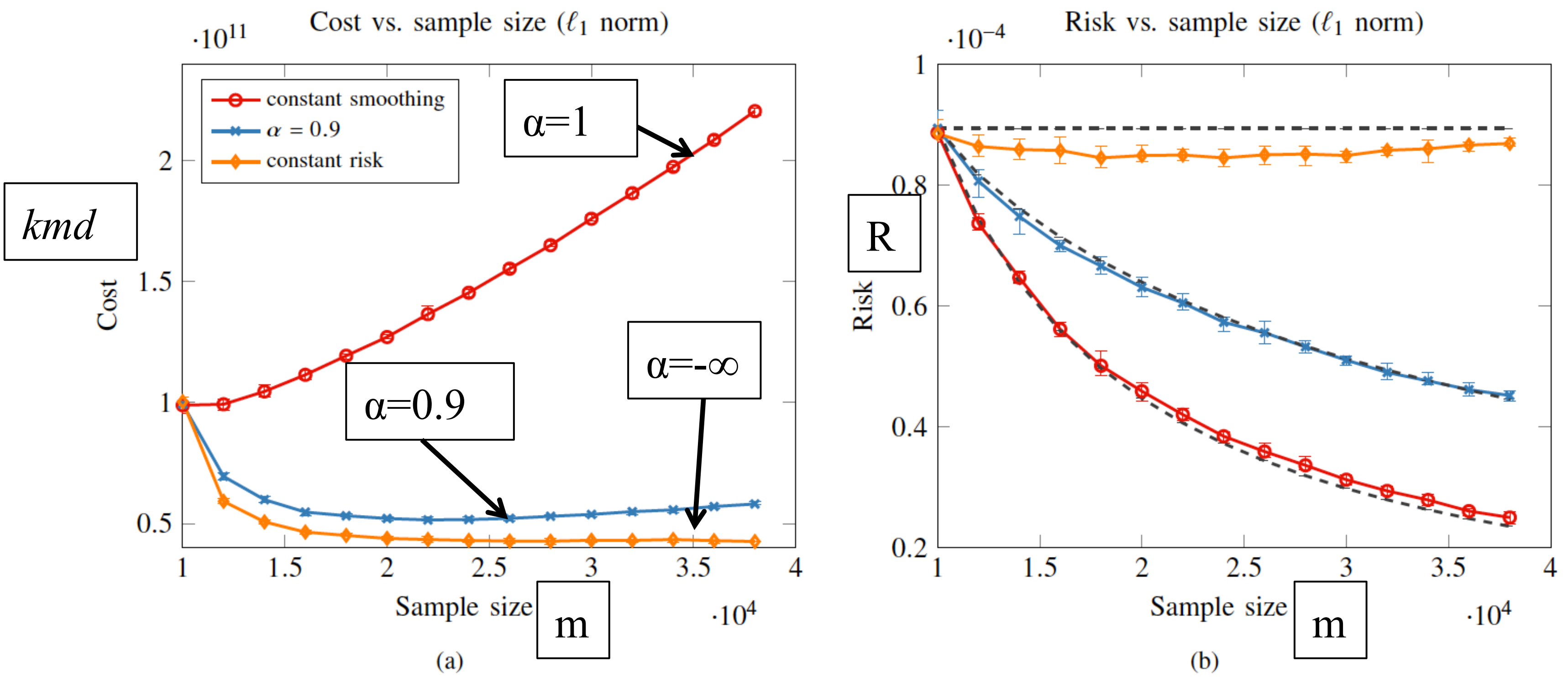
R

m

(b)

Fig. 2. **Sparse vector regression experiment.** The panels show (a) the average computational cost and (b) the estimated statistical risk over 10 random trials of the dual-smoothed sparse vector regression problem with ambient dimension $d = 40\,000$, normalized sparsity $\rho = 5\%$, and noise level $\sigma = 0.01$ for various sample sizes $m$. The red curve (circles) represents using a fixed smoothing parameter $\mu = 0.1$, the orange curve (diamonds) results from adjusting the smoothing parameter $\mu$ to maintain the baseline risk, and the blue curve (crosses) uses the balanced scheme (12) with scaling parameter $\alpha = 0.9$. For all schemes, the baseline smoothing parameter $\overline{\mu} = 0.1$, and the baseline sample size $\overline{m} = 10\,000$. The error bars indicate the minimum and maximum observed values. The dashed black lines show the predicted risk based on Proposition VI.1 and Fact III.3.

# LASSO



**α = 1**: $\mu$ is constant; δ is constant; **R falls; cost increases**
**α = -∞**: $\mu$ increases; δ increases; **R is constant; cost decreases**
**α in (-∞,1)**: ***balance!***

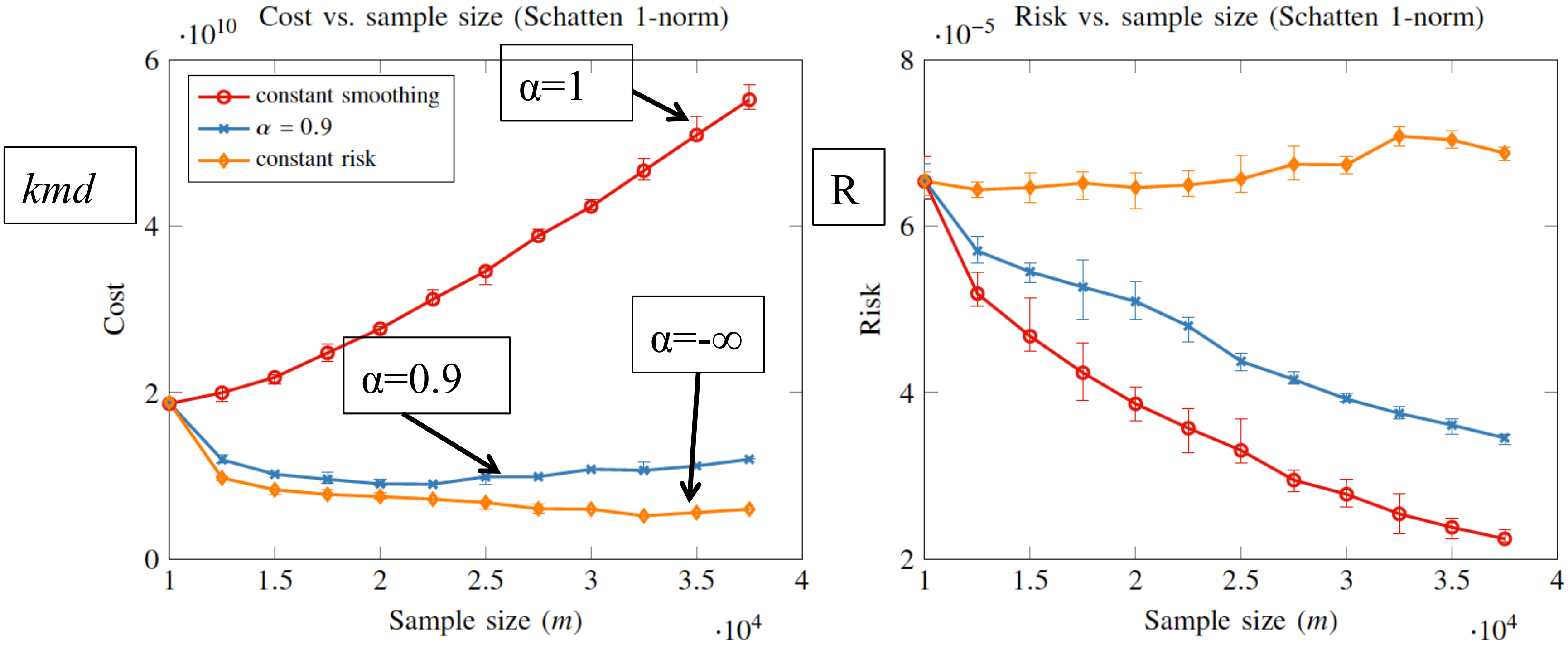
# Low-Rank Matrix Regression

$$\widehat{\boldsymbol{x}}_{\mu} := \arg\min_{\boldsymbol{x}} f_{\mu}(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\| \le \epsilon.$$

$$f_{\mu}(\boldsymbol{X}) = \|\boldsymbol{X}\|_{*} + \frac{\mu}{2}\|\boldsymbol{X}\|_{\mathrm{F}}^{2}$$

$$\psi(\rho) := \inf_{0 \le \tau \le 2} \Big\{ \rho + (1-\rho)\Big[\rho\left(1 + \tau^2(1 + \mu\|\boldsymbol{X}\|)^2\right) + \frac{(1-\rho)}{12\pi}\left[24(1+\tau^2)\cos^{-1}(\tau/2) - \tau(26+\tau^2)\sqrt{4-\tau^2}\right]\Big]\Big\}.$$

Again, we have some complicated looking upper bound for δ.



(a)

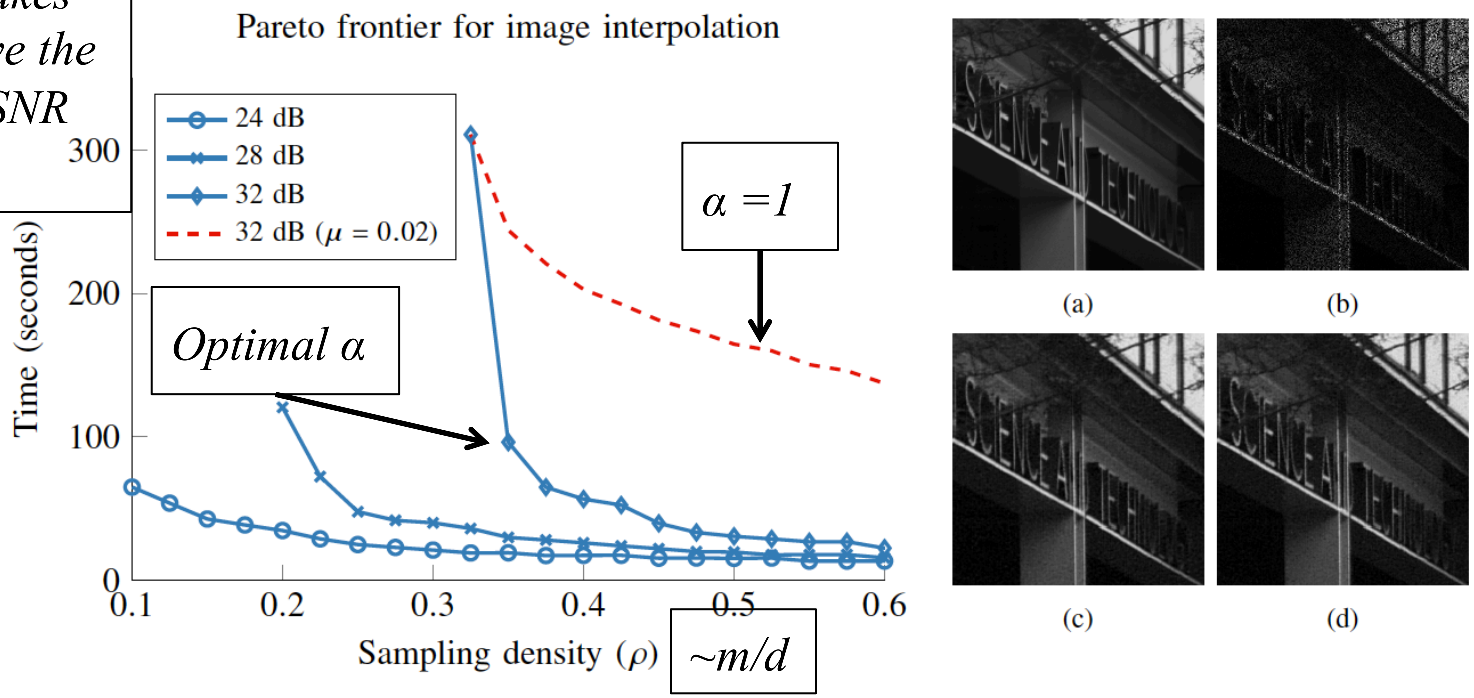
(b)

Fig. 3. **Low-rank matrix regression experiment.** The panels show (a) the average computational cost and (b) the estimated statistical risk over 10 random trials of the dual-smoothed low-rank matrix regression problem with ambient dimension $d = 200 \times 200$, normalized rank $\rho = 5\%$, and noise level $\sigma = 0.01$ for various sample sizes $m$. The red curve (circles) represents using a fixed smoothing parameter $\mu = 0.1$, the orange curve (diamonds) results from adjusting the smoothing parameter $\mu$ to maintain the baseline risk, and the blue curve (crosses) uses the balanced scheme (12) with scaling parameter $\alpha = 0.9$. For all schemes, the baseline smoothing parameter $\overline{\mu} = 0.1$, and the baseline sample size $\overline{m} = 10\,000$. The error bars indicate the minimum and maximum observed values.

# Image Interpolation

$$\begin{aligned} &\text{minimize} \quad \|\mathcal{W}(X)\|_{\ell_1} + \frac{\mu}{2}\|\mathcal{W}(X)\|^2 \\ &\text{subject to} \quad \mathcal{A}(X) = \boldsymbol{b}, \end{aligned}$$

*W* is 2D discrete cosine transformation (DCT) operator.
We expect this to be sparse for natural images.



*Time it takes to achieve the target PSNR*

*Optimal α*

*α =1*

*~m/d*

Fig. 4. **Image interpolation.** The graph shows the observed Pareto frontier in our image interpolation experiment where we treat the sampling density $\rho$ and computational time as the two resources that we trade off. The solid blue lines give the Pareto frontiers achieved by aggressively smoothing the problem as we increase the sampling density $\rho$. These frontiers correspond to three different accuracy levels of the reconstructed images given as a peak signal-to-noise-ratio (PSNR). The dashed red line shows the frontier achieved for 32 dB PSNR accuracy with a fixed smoothing parameter $\mu = 0.02$ as sampling density $\rho$ increases. Our aggressive smoothing outperforms the constant smoothing by a large margin. The grid of images shows $450 \times 450$ pixel patches of: (a) the original image, (b) the original image subsampled at $\rho = 40\%$, (c) the reconstructed image with $\rho = 40\%$ and $\mu = 0.02$ (32.1 dB PSNR), and (d) the reconstructed image with $\rho = 40\%$ and $\mu = 0.32$ (32.2 dB PSNR). The shown reconstructions are of the same quality despite the differing values of $\mu$.

Peak signal to noise ratio $\mathrm{PSNR}(\boldsymbol{X}_k) = 10 \cdot \log_{10}\left(\frac{d_1 d_2}{\|\boldsymbol{X}_k - \boldsymbol{X}^\natural\|_{\mathrm{F}}}\right),$

# couple of comments…

- Why not use the actual time taken instead of *kmd* in the numerical simulations? Or does that even matter?
- To use phase transition theorem, measurement matrix A had to have orthonormal rows; is that restriction costly in any way?
- They mention the possibility of tuning $\mu$ to meet target R. That would be really interesting to see.

# References

[1] ‘‘Designing Statistical Estimators That Balance Sample Size, Risk, and Computational Cost,‘‘ J. J. Bruer, J. A. Tropp, V. Cevher, and S. Becker, *IEEE Journal of Selected Topics in Signal Processing*, 2015.

[2] S. Shalev-Shwartz and N. Srebro, “SVM optimization: inverse dependence on training set size,” in Proc. 25th Annu. Int. Conf. Machine Learning (ICML 2008), pp. 928–935, ACM, 2008.

[3] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, “Living on the edge: A geometric theory of phase transitions in convex optimization,” Information and Inference, vol. to appear, 2014.

[4] S. Oymak and B. Hassibi, “Sharp MSE Bounds for Proximal Denoising,” arXiv, 2013, 1305.2714v5.

[5] A. Auslender and M. Teboulle, “Interior gradient and proximal methods for convex and conic optimization,” SIAM J. Optim., vol. 16, no. 3, pp. 697–725, 2006.

[6] S. R. Becker, E. J. Candès, and M. C. Grant, “Templates for convex cone problems with applications to sparse signal recovery,” Math. Program. Comput., vol. 3, no. 3, pp. 165–218, 2011.