# Compressed Sensing over $\ell_p$-balls: Minimax Mean Square Error

David Donoho, Iain Johnstone, Arian Maleki and Andrea Montanari

March 23, 2011

## Abstract

We consider the compressed sensing problem, where the object $x_0 \in \mathbf{R}^N$ is to be recovered from incomplete measurements $y = Ax_0 + z$; here the sensing matrix $A$ is an $n \times N$ random matrix with iid Gaussian entries and $n < N$. A popular method of sparsity-promoting reconstruction is $\ell^1$-penalized least-squares reconstruction (aka LASSO, Basis Pursuit).

It is currently popular to consider the strict sparsity model, where the object $x_0$ is nonzero in only a small fraction of entries. In this paper, we instead consider the much more broadly applicable $\ell_p$-sparsity model, where $x_0$ is sparse in the sense of having $\ell_p$ norm bounded by $\xi \cdot N^{1/p}$ for some fixed $0 < p \leq 1$ and $\xi > 0$.

We study an asymptotic regime in which $n$ and $N$ both tend to infinity with limiting ratio $n/N = \delta \in (0, 1)$, both in the noisy ($z \neq 0$) and noiseless ($z = 0$) cases. Under weak assumptions on $x_0$, we are able to precisely evaluate the worst-case asymptotic minimax mean-squared reconstruction error (AMSE) for $\ell^1$ penalized least-squares: min over penalization parameters, max over $\ell_p$-sparse objects $x_0$. We exhibit the asymptotically least-favorable object (hardest sparse signal to recover) and the maximin penalization.

In the case where $n/N$ tends to zero slowly – i.e. extreme undersampling – our formulas (normalized for comparison) say that the minimax AMSE of $\ell_1$ penalized least-squares is asymptotic to $\xi^2 \cdot (\frac{2 \log(N/n)}{n})^{2/p-1} \cdot (1 + o(1))$. Thus we have not only the rate but also the constant factor on the AMSE; and the maximin penalty factor needed to attain this performance is also precisely specified. Other similarly precise calculations are showcased.

Our explicit formulas unexpectedly involve quantities appearing classically in statistical decision theory. Occurring in the present setting, they reflect a deeper connection between penalized $\ell^1$ minimization and scalar soft thresholding. This connection, which follows from earlier work of the authors and collaborators on the AMP iterative thresholding algorithm, is carefully explained.

Our approach also gives precise results under weak-$\ell_p$ ball coefficient constraints, as we show here.

**Key Words:** Approximate Message Passing. Lasso. Basis Pursuit. Minimax Risk over Nearly-Black Objects. Minimax Risk of Soft Thresholding.

# 1 Introduction

In the compressed sensing problem, we are given a collection of noisy, linear measurements of an unknown vector $x_0$

$$y = Ax_0 + z, \tag{1.1}$$

Here the measurement matrix $A$ has dimensions $n$ by $N$, $n < N$, the $N$-vector $x_0$ is the object we wish to recover and the noise $z \sim \mathsf{N}(0, \sigma^2 \mathbf{I})$. Both $y$ and $A$ are known, both $x_0$ and $z$ are unknown, and we seek an approximation to $x_0$.

Since the equations are underdetermined and noisy, it seems hopeless to recover $x_0$ in general, but in compressed sensing one also assumes that the object is *sparse*. In a number of recent papers, the sparsity assumption is formalized by requiring $x_0$ to have at most $k$ nonzero entries. This *k-sparse model* leads to a simpler analysis, but is highly idealized, and does not cover situations where a few dominant entries are scattered among many small but slightly nonzero entries. For such situations, [Don06a] proposed to measure sparsity by membership in $\ell_p$ balls $0 < p \le 1$, namely to consider the situation where the $\ell_p$-norm[1] of $x_0$ is bounded as

$$\|x_0\|_p^p \equiv \sum_{i=1}^{p} |x_{0,i}|^p \le N\xi^p, \tag{1.2}$$

for some constraint parameter $\xi$. Here, as $p \to 0$, we recover the $k$-sparse case (aka $\ell_0$ constraint).

Much more is known today about behavior of reconstruction algorithms under the $k$-sparse model than in the more realistic $\ell_p$ balls model. In some sense the $k$-sparse model has been more amenable to precise analysis. In the noiseless setting, precise asymptotic formulas are now known for the sparsity level $k$ at which $\ell_1$ minimization fails to correctly recover the object $x_0$ [Don06b, DT05, DT10]. In the noisy setting, precise asymptotic formulas are now known for the worst-case asymptotic mean-squared error of reconstruction by $\ell_1$-penalized $\ell_2$ minimization [DMM10, BM11]. By comparison, existing results for the $\ell_p$ balls model are mainly qualitative estimates, i.e. bounds that capture the correct scaling with the problem dimensions but involve loose or unspecified multiplicative coefficients. We refer to Section 10.2 for a brief overview of this line of work, and a comparison with our results.

We believe our paper brings the state of knowledge about the $\ell_p$-ball sparsity model to the same level of precision as for the $k$-sparse model. We consider here the high-dimensional setting $N, n \to \infty$ with matrices $A$ having iid Gaussian entries. We treat both the noisy and noiseless cases in a unified formalism and provide precise expressions, including constants, describing the worst-case large-system behavior of mean-squared error for optimally-tuned $\ell^1$-penalized reconstructions. Because our expressions are precise, they deserve close scrutiny; as we show here, this attention is rewarded with surprising insights, such as the equivalence of undersampling with adding additional noise. Less precise methods could not provide such insights.

The rest of this introduction reviews the results obtained through our method.

---

[1]Throughout this paper we will accept the abuse of terminology of calling $\| \cdot \|_p$ a 'norm', although it is not a norm for $p < 1$.

## 1.1 Problem formulation; Preview of Main Results

Our main results concern $\ell^1$-penalized least-squares reconstruction with penalization parameter $\lambda$.

$$\widehat{x}_\lambda \equiv \arg\min_x \left\{ \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 \right\}. \tag{1.3}$$

This reconstruction rule became popular under the names of LASSO [Tib96] or Basis Pursuit DeNoising [CD95]. Our analysis involves a large-system limit, which was effectively also used in [DMM09, DMM10, BM11]. We introduce some convenient terminology:

**Definition 1.1.** *A* **problem instance** $I_{n,N}$ *is a triple* $I_{n,N} = (x_0^{(N)}, z^{(n)}, A^{(n,N)})$ *consisting of an object* $x_0^{(N)}$ *to recover, a noise vector* $z^{(n)}$, *and a measurement matrix* $A$. *A* **sequence of instances** $\mathsf{S} = (I_{n,N})$ *is an infinite sequence of such problem instances.*

At this level of generality, a sequence of instances is nearly arbitrary. We now make specific assumptions on the members of each triple. HEre and below $\mathbb{I}(\mathcal{P})$ is the indicator function on property $\mathcal{P}$.

**Definition 1.2.** • **Object** $\ell_p$ **sparsity constraint.** *A sequence* $\mathbf{x}_0 = (x_0^{(N)})$ *belongs to* $\mathcal{X}_p(\xi)$ *if (i)* $N^{-1}\|x_0^{(N)}\|_p^p \leq \xi^p$, *for amm M; and (ii) There exists a sequence* $B = \{B_M\}_{M\geq 0}$ *such that* $B_M \to 0$, *and for every* $N$, $\sum_{i=1}^N (x_{0,i}^{(N)})^2 \mathbb{I}(|x_{0,i}^{(N)}| \geq M) \leq B_M N$.

- **Noise power constraint.** *A sequence* $\mathbf{z} = (z^{(n)})_n$ *belongs to* $\mathcal{Z}^2(\sigma)$ *if* $n^{-1}\|z^{(n)}\|_2^2 \to \sigma^2$.

- **Gaussian Measurement matrix.** $A^{(n,N)} \sim \text{GAUSS}(n,N)$ *is an* $n \times N$ *random matrix with entries drawn iid from the* $\mathsf{N}(0, \frac{1}{n})$ *distribution.*

- *The* **Standard** $\ell_p$ **Problem Suite** $\mathcal{S}_p(\delta, \xi, \sigma)$ *is the collection of sequences of instances* $\mathsf{S} = \{(x_0^{(N)}, z^{(n)}, A^{(n,N)})\}_{n,N}$ *where*
  *(i)* $n/N \to \delta$,
  *(ii)* $\mathbf{x}_0 \in \mathcal{X}_p(\xi)$,
  *(iii)* $\mathbf{z} \in \mathcal{Z}^2(\sigma)$, *and*
  *(iv) each* $A^{(n,N)}$ *is sampled from the Gaussian ensemble* $\text{GAUSS}(n,N)$.

The uniform intergrability condition $\sum_{i=1}^N (x_{0,i}^{(N)})^2 \mathbb{I}(|x_{0,i}^{(N)}| \geq M) \leq B_M N$ essentially requires that the $\ell_2$ norm of $x_0^{(N)}$ is not dominated by a small subset of entries. As we discuss below, it is a fairly weak condition and most likely can be removed because the least-favorable vectors $x_0$ turn out to have all non-zero entries of the same magnitude. Finally notice that uniform integrability is implied by following: there exist $q > 2$, $B < \infty$ such that $\|x_0^{(N)}\|_q^q \leq NB$ for all $N$.

The fraction $\delta = n/N$ measures the incompleteness of the underlying systems of equations, with $\delta$ near 1 meaning $n \approx N$ and so nearly complete sampling, and $\delta$ near 0 meaning $n \ll N$ and so highly incomplete sampling.

Note in particular: the estimand $\mathbf{x}$ and the noise $\mathbf{z}$ are deterministic sequences of objects, while the matrix $A$ is random. In particular, while it may seem natural to pick the noise to be random, that is not necessary, and in fact plays no role in our results.

Also let AMSE($\lambda$; S) denote the asymptotic per-coordinate mean-squared error of the LASSO reconstruction with penalty parameter $\lambda$, for the sequence of problem instances S :

$$\text{AMSE}(\lambda, \mathsf{S}) = \limsup \frac{1}{N} \mathbb{E}\big\{\|\widehat{x}_\lambda^{(N)} - x_0^{(N)}\|^2\big\}. \tag{1.4}$$

Here $\widehat{x}_\lambda^{(N)}$ denotes the LASSO estimator, and $x_0^{(N)}$ the estimand, on problem instances of size[2] $N$. Moreover the limsup is taken as $n, N \to \infty; n \sim \delta N$. Although in general this quantity need not be well defined, our results imply that, if the sequence of instances S is taken from the standard problem suite, this quantity is bounded.

Now the AMSE depends on both $\lambda$, the penalization parameter, and $\mathbf{x}$, the sequence of objects to recover. As in traditional statistical decision theory, we may view the AMSE as the payoff function of a game against Nature, where Nature chooses the object sequence $\mathbf{x}$ and the researcher chooses the threshold parameter $\lambda$. In this paper, Nature is allowed to pick only sparse objects $x_0^{(N)}$ obeying the constraint $N^{-1}\|x_0^{(N)}\|_p^p \leq \xi^p$.

In the case of noiseless information, $y = Ax_0$ (so $z = 0$), this game has a saddlepoint, and Theorem 4.1 gives a precise evaluation of the minimax AMSE:

$$\sup_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,0)} \inf_\lambda \text{AMSE}(\lambda, \mathsf{S}) = \frac{\delta\xi^2}{M_p^{-1}(\delta)^2}. \tag{1.5}$$

The maximin on the left side is the payoff of a zero-sum game.

The function on the right side, $M_p(\,\cdot\,)$ is displayed in Figure 1. It evaluates the minimax MSE in a classical and much discussed problem of statistical decision theory: soft threshold estimation of random means $X$ satisfying the moment constraint $\mathbb{E}\{|X|^p\} \leq \xi^p$ from noisy data $X + \mathsf{N}(0, 1)$. This problem was studied in [DJ94], and detailed information is known about $M_p$; see Section 2 for a review.

In the noisy case, $\sigma > 0$, we have the same setup as before, only now the AMSE will of course be larger. Theorem 5.1 gives the minimax AMSE precisely:

$$\sup_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,\sigma)} \inf_\lambda \text{AMSE}(\lambda, \mathsf{S}) = \sigma^2 \cdot m_p^*(\delta, \xi/\sigma), \tag{1.6}$$

where $m_p^* = m_p^*(\delta, \xi)$ is defined as the unique positive solution of the equation

$$\frac{m}{1 + m/\delta} = M_p\left(\frac{\xi}{(1 + m/\delta)^{1/2}}\right). \tag{1.7}$$

Again, the precise formula involves $M_p(\,\cdot\,)$, a classical quantity in statistical decision theory. See Figure 8 for a display of the minimax AMSE as a function of $p$ and $\xi$.

Our results include several other precise formulas; our approach is able to evaluate a number of operationally important quantities

- The *least-favorable* object, ie. the sparse estimand $x_0$ which causes maximal difficulty for the LASSO; Eqs (4.4), (5.5), (6.6).

---

[2]It would be more notationally correct to write $\widehat{x}_\lambda^{(N,n)}$ since the full problem size involves both $n$ and $N$, but we ordinarily have in mind a specific value $\delta \sim n/N$, hence $n$ is not really free to vary independent of $N$.

- The *maximin tuning*, the actual choice of penalization which minimizes the AMSE when Nature chooses the least-favorable distribution; Eqs (4.3), (5.6), (6.16).

- Various operating characteristics, including the AMSE of reconstruction, and the limiting $\ell_p$ norms of the reconstruction.

Various figures and tables present precise calculations which one can make using the results of this paper. Figure 5 shows the Minimax AMSE as a function of $\delta > 0$, for the noiseless case $z = 0$ with fixed $\xi = 1$, while Figure 8 gives the minimax AMSE as a function of $\xi$ for fixed $\delta = 1/4$, for the noisy case where the mean-square value of $z$ is $\sigma^2$.

## 1.2    Novel Interpretations

Our precise formulas provide not only accurate numerical information, but also rather surprising insights. The appearance of the classical quantity $M_p$ in these formulas tells us that a *noiseless* compressed sensing problem, with *nonsquare* sensing matrix $A$ having $n < N$ is explicitly connected with the MSE in a very simple *noisy* problem where $n = N$, $A$ is *square* – in fact, the identity(!) – cf. Eq. (1.5). On the other hand, a *noisy* compressed sensing problem with $n < N$ and so $A$ nonsquare is explicitly connected with a seemingly trivial problem, where $n = N$ and $A$ is the identity, but the noise level is *different* than in the compressed sensing problem – in fact *higher* – cf. Eqs. (1.6), (1.7). Conclusion:

> **Slogan:** *In both the noisy and noiseless cases: undersampling is effectively equivalent to adding noise to complete observations.*[3]

While [DTDS06] and [LDSP08] formulate heuristics and provided empirical evidence about this connection, the results here (and in the companion papers [DMM09, DMM10]) provide the only theoretical derivation of such a connection.

Established research tools for understanding compressed sensing - for example estimates based on the restricted isometry property [CT05, CRT06] - provide upper bounds on the mean square error but do not allow one to suspect that such striking connections hold. In fact we use a very different approach from the usual compressed sensing literature. Our methods join ideas from belief propagation message passing in information theory, and minimax decision theory in mathematical statistics.

## 1.3    Complements and Extensions

### 1.3.1    Weak $\ell_p$

Section 6 develops analogous results for compressed sensing in the weak-$\ell_p$ balls model, where the object obeys a *weak-$\ell_p$* rather than an $\ell_p$ constraint. Weak-$\ell_p$ balls are relevant models for natural images and hence our results have applications in image reconstruction, as we describe in Section 9.

---

[3]The formal equivalence of undersampling to simply adding noise is quite striking. It reminds us of ideas from the so-called comparison of experiments in traditional statistical decision theory.

### 1.3.2   Reformulation of $\ell_p$ Balls

Our normalization of the error measure and of $\ell_p$ balls are somewhat different than what has been called the $\ell_p$ case in earlier literature. We also impose a tightness condition not present in earlier work. In exchange, we get precise results. For calibration of these results see Section 7. From the practical point of view of obtaining *accurate predictions about the behavior of real systems*, the present model has significant advantages. For more detail, see Section 10.

## 2   Minimax Mean Squared Error of Soft Thresholding

Consider a signal $x_0 \in \mathbb{R}^N$, and suppose that it satsifies $x_0$ satisfies the $\ell_2$-normalization $N^{-1}\|x_0\|_p^p \approx 1$ but also the $\ell_p$-constraint $\|x_0\|_p^p \leq N \cdot \xi^p$, for small $\xi$ and $0 < p < 2$. To see that this is a sparsity constraint, note that a typical 'dense' sequence, such as an iid Gaussian sequence, cannot obey such a constraint for large $N$; in effect, smallness of $\xi$ rules out sequences which have too many significantly nonzero values.

If we observed such a sparse sequence in additive Gaussian noise $y = x_0 + z$, where $z \sim_{iid} \mathsf{N}(0,1)$, it is well-known that we could approximately recover the vector by simple thresholding – effectively, zeroing out the entries which are already close to zero. Consider the soft-thresholding nonlinearity $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$. Given an observation $y \in \mathbb{R}$ and a 'threshold level' $\tau \in \mathbb{R}_+$, soft thresholding acts on a scalar as follows

$$
\eta(y;\tau) = \begin{cases} y - \tau & \text{if } y \geq \tau, \\ 0 & \text{if } -\theta < y < \tau, \\ y + \tau & \text{if } y \leq -\tau. \end{cases} \tag{2.1}
$$

We apply it to a vector $y$ coordinatewise and get the estimate $\hat{x} = \eta(y;\tau)$.

To analyze this procedure we can work in terms of scalar random variables. The empirical distribution of $x_0$ is defined as

$$
\nu_{x_0,N} \equiv \frac{1}{N} \sum_{i=1}^{N} \delta_{x_{0,i}}. \tag{2.2}
$$

Define the random variables $X \sim \nu_{x_0,N}$ and $Z \sim \mathsf{N}(0,1)$, with $X$ and $Z$ mutually independent. We have the isometry:

$$
N^{-1}\mathbb{E}\|\hat{x} - x_0\|_2^2 = \mathbb{E}_{\nu_{x_0}}\left\{ \left[\eta(X + Z;\tau) - X\right]^2 \right\}.
$$

Hence, to analyze the behavior of thresholding under sparsity constraints, we can shift attention from sequences in $\mathbb{R}^N$ to distributions.

So define the class of 'sparse' probability distributions over $\mathbb{R}$:

$$
\mathcal{F}_p(\xi) \equiv \left\{ \nu \in \mathcal{P}(\mathbb{R}) : \nu(|X|^p) \leq \xi^p \right\}, \tag{2.3}
$$

where $\mathcal{P}(\mathbb{R})$ denotes the space of probability measures over the real line. Then $x_0$ satisfies the $\ell_p$-constraint $\|x_0\|_p^p \leq N \cdot \xi^p$ if and only if $\nu_{x_0} \in \mathcal{F}_p(\xi)$.

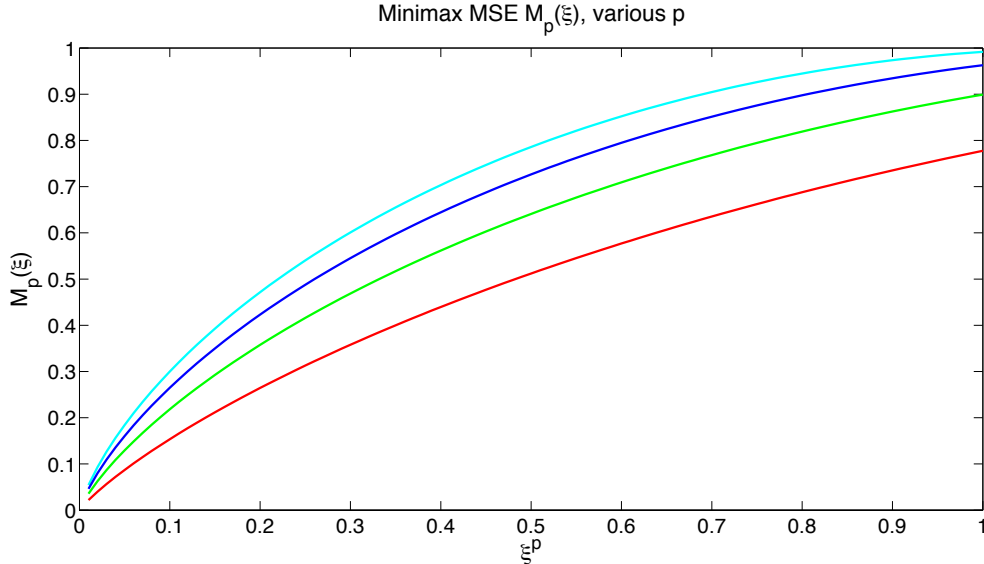The central quantity for our formulae (1.5), (1.6) is the minimax mean square error $M_p(\xi)$ defined now:

Figure 1: Minimax soft thresholding risk, $M_p(\xi)$, various $p$. Vertical axis: worst case MSE over $\mathcal{F}_p(\xi)$. Horizontal axis: $\xi^p$. Red, green, blue, aqua curves correspond to $p = 0.1, 0.25, 0.50, 1.00$.

**Definition 2.1.** *The* minimax mean squared error *of soft thresholding is defined by:*

$$M_p(\xi) = \inf_{\tau \in \mathbb{R}_+} \sup_{\nu \in \mathcal{F}_p(\xi)} \mathbb{E}\left\{\left[\eta(X + Z; \tau) - X\right]^2\right\}, \tag{2.4}$$

*where expectation on the right hand side is taken with respect to $X \sim \nu$ and $Z \sim \mathsf{N}(0,1)$ mutually independent.*

This quantity has been carefully studied in [DJ94], particularly in the asymptotic regime $\xi \to 0$. Figure 1 displays its behavior as a function of $\xi$ for several different values of $p$.

The quantity (2.4) can be viewed as the value of a game against Nature, where the statistician chooses the threshold $\tau$, Nature chooses the distribution $\nu$, and the statistician pays Nature an amount equal to the MSE. We use the following notation for the MSE of soft thresholding, given a noise level $\sigma$, a signal distribution $\nu$ and a threshold level $\tau$:

$$\mathsf{mse}(\sigma^2; \nu, \tau) \equiv \mathbb{E}\left\{\left[\eta(X + \sigma Z; \tau \sigma) - X\right]^2\right\}, \tag{2.5}$$

where, again, expectation is with respect to $X \sim \nu$ and $Z \sim \mathsf{N}(0,1)$ independent. Hence the quantity on the right hand side of Eq. (2.4) –the game payoff– is just $\mathsf{mse}(1; \nu, \tau)$.

Evaluating the supremum in Eq. (2.4) might at first appear hopeless. In reality the computation can be done rather explicitly using the following result.

**Lemma 2.1.** *The least-favorable distribution $\nu_{p,\xi}$, i.e. the distribution forcing attainment of the worst-case MSE, is supported on 3 points. Explicitly, consider the 3-point mixture distribution*

$$\nu_{\varepsilon,\mu} = (1 - \varepsilon)\delta_0 + \frac{\varepsilon}{2}\delta_\mu + \frac{\varepsilon}{2}\delta_{-\mu}. \tag{2.6}$$

7

*Then the least-favorable distribution $\nu_{p,\xi}$ is the 3-point mixture $\nu_{\varepsilon_p(\xi),\mu_p(\xi)}$ for specific values $\varepsilon_p(\xi), \mu_p(\xi)$.*

In fact it seems the minimax problem in Eq. (2.4) has a saddlepoint, i.e. a pair $(\nu_{p,\xi}, \tau_p(\xi)) \in \mathcal{P}(\mathbb{R}) \times \mathbb{R}_+$, such that

$$\mathsf{mse}(1; \nu_{p,\xi}, \tau) \geq \mathsf{mse}(1; \nu_{p,\xi}, \tau_p(\xi)) \geq \mathsf{mse}(1; \nu, \tau_p(\xi)) \qquad \forall \tau > 0, \ \forall \nu \in \mathcal{F}_p(\xi), \qquad (2.7)$$

but we do not need or prove this fact here. The MSE is readily evaluated for 3-point distribution, yielding

$$\mathsf{mse}(1; \nu_{\varepsilon,\mu}, \tau) = (1 - \varepsilon)\big\{2(1 + \tau^2)\Phi(-\tau) - 2\tau\phi(\tau)\big\} \qquad\qquad\qquad (2.8)$$
$$+ \varepsilon\big\{\mu^2 + (1 + \tau^2 - \mu^2)[\Phi(-\mu - \tau) + \Phi(\mu - \tau)] + (\mu - \tau)\phi(\mu + \tau) - (\mu + \tau)\phi(-\mu + \tau)\big\}.$$

Here and below, $\phi(z) \equiv e^{-z^2/2}/\sqrt{2\pi}$ is the standard Gaussian density and $\Phi(x) \equiv \int_{-\infty}^{x} \phi(z)\,\mathrm{d}z$ is the Gaussian distribution function. Further, it is easy to check that the MSE is maximized when the $\ell_p$ constraint is saturated, i.e. for

$$\varepsilon\mu^p = \xi^p. \qquad\qquad\qquad (2.9)$$

Therefore one is left with the task of maximizing the right-hand side of Eq. (2.8) with respect to $\varepsilon$ (for $\mu = \xi\varepsilon^{-1/p}$) and minimizing it with respect to $\tau$. This can be done quite easily numerically for any given $\xi > 0$, yielding the values of $\tau_p(\xi)$, $\mu_p(\xi)$ and $\varepsilon_p(\xi)$ plotted in Fig. 2. The minimax property is illustrated in Fig. 3.

Important below will be the inverse function

$$M_p^{-1}(m) = \inf\big\{\,\xi \in [0, \infty) \,:\, M_p(\xi) \geq m\,\big\}, \qquad\qquad\qquad (2.10)$$

defined for $m \in (0, 1)$, and depicted in Figure 4. The well-definedness of this function follows from the next Lemma.

**Lemma 2.2.** *The function $\xi \mapsto M_p(\xi)$ is continuous and strictly increasing for $\xi \in (0, \infty)$, with $\lim_{\xi \to 0} M_p(\xi) = 0$, and $\lim_{\xi \to \infty} M_p(\xi) = 1$.*

*Proof.* Let $\mathsf{mse}_0(\mu, \tau) \equiv \mathbb{E}\{[\eta(\mu + Z; \tau) - \mu]^2\}$ for $Z \sim \mathsf{N}(0, 1)$, so that $\mathsf{mse}(1; \tau, \nu) = \int \mathsf{mse}_0(\mu, \tau)\,\nu(\mathrm{d}\mu)$. Since $\mathsf{mse}_0(\mu, \tau) = \mathsf{mse}_0(-\mu, \tau)$ in this formula we can assume without loss of generality that $\nu(\cdot)$ is supported on $\mathbb{R}_+$.

To show strict monotonicity, fix $\xi \leq \xi'$, let $\tau' = \tau_p(\xi')$ be the minimax threshold for $\mathcal{F}_p(\xi')$, and let $\nu_\xi = \nu_{p,\xi}$ be the least favorable prior for $\mathcal{F}_p(\xi)$. Let $\nu' = S_{\xi'/\xi}\nu_\xi$ be the measure in $\mathcal{F}_p(\xi')$ obtained by scaling $\nu_\xi$ up by a factor $\xi'/\xi$ (explicitly, for a measurable set $C$, $\nu'(C) = \nu_\xi((\xi'/\xi)C)$). Since $\nu_\xi \neq \delta_0$, strict monotonicity of $\mu \to \mathsf{mse}_0(\mu, \tau)$ (e.g. [DJ94, eq. A2.8]) shows that $\mathsf{mse}(1; \tau', \nu_\xi) < \mathsf{mse}(1; \tau', \nu')$. Consequently

$$M_p(\xi) \leq \mathsf{mse}(1; \tau', \nu_\xi) < \mathsf{mse}(1; \tau', \nu') \leq \sup_{\nu \in \mathcal{F}_p(\xi')} \mathsf{mse}(1; \tau', \nu) = M_p(\xi').$$

We verify that $t \to M_p(t^{1/p})$ is concave in $t$: combined with strict monotonicity, we can then conclude that $M_p(\xi)$ is continuous. Indeed, the map $\nu \to \mathsf{mse}(1; \tau, \nu)$ is linear in $\nu$ and
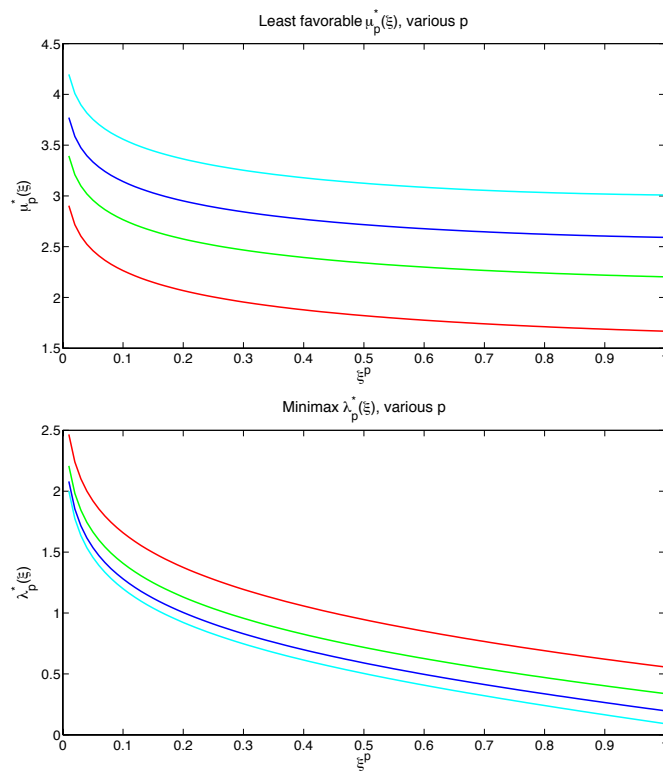
Figure 2: Least-favorable $\mu$ (upper frame) and corresponding minimax threshold $\tau$ (lower frame). Horizontal axes: $\xi^p$. Red, green, blue, aqua curves correspond to $p = 0.1, 0.25, 0.50, 1.00$.
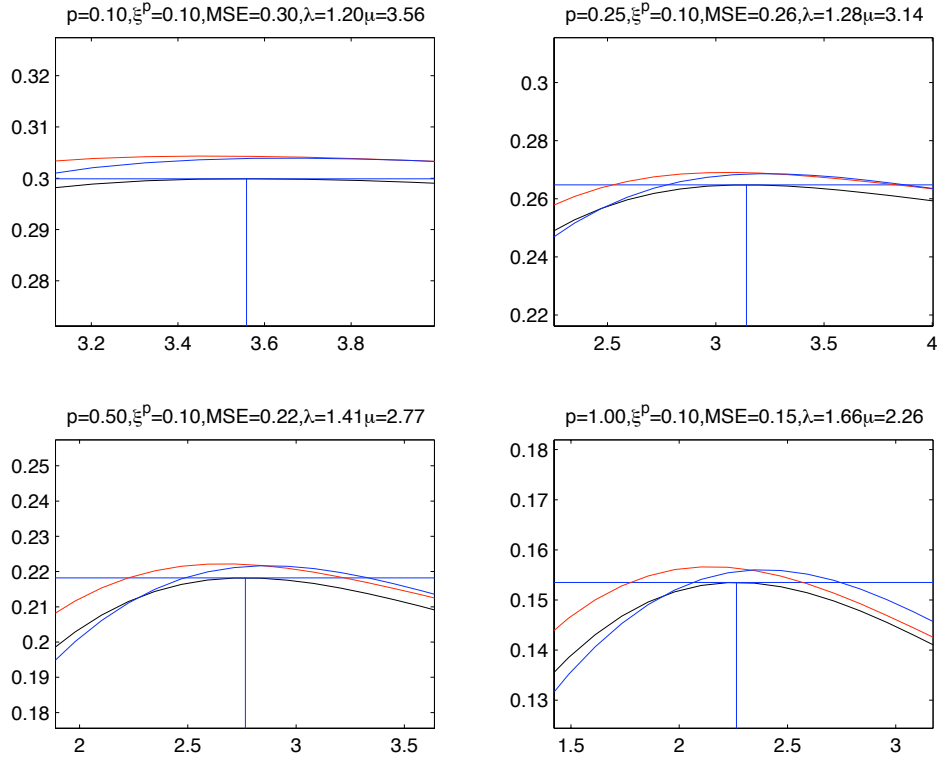
Figure 3: Saddlepoint property of Minimax $\tau_p(\xi)$, $\xi^p = 1/10$, various $p$. Vertical Axis: MSE at $F_{\varepsilon,\mu}$. Horizontal Axis $\mu$. Vertical Blue line: least-favorable $\mu$, $\mu_p(\xi)$. Horizontal Blue Line: Minimax MSE $M_p(\xi)$. At each value of $\mu$, Black curve displays corresponding MSE of soft thresholding with threshold at the minimax threshold value $\tau_p(\xi)$ , under the distribution $F_{\varepsilon,\mu}$ with $\varepsilon\mu^p = \xi^p$. The other two curves are for $\tau$ 10 percent higher and 10 percent lower than the minimax value. In each case, the black curve (associated with minimax $\tau$), stays below the horizontal line, while the red and blue curves cross above it, illustrating the saddlepoint relation.
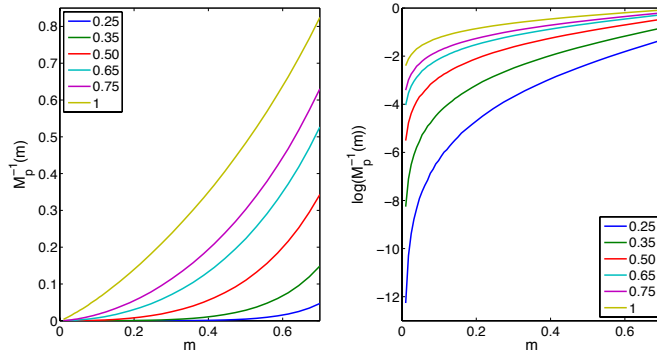
Figure 4: Inverse function $M_p^{-1}(m)$. Horizontal axis: $m$, desired minimax mean square error $m$. Vertical axis: left-hand plot: $\xi = M_p^{-1}(m)$, the radius of ball that attains it. right-hand plot: $\log(\xi)$. Colored curves correspond to various choices of $p$.

so $\mathsf{mse}_*(\nu) = \inf_\tau \mathsf{mse}(1; \tau, \nu)$ is concave in $\nu$. Hence $M_p(t^{1/p}) = \sup\{\mathsf{mse}_*(\nu) \; : \; \nu(|X|^p) \le t\}$ is also concave.

That $\lim_{\xi \to 0} M_p(\xi) = 0$ is shown in [DJ94], compare Lemma 2.3 below. For large $\xi$, observe that

$$1 \ge M_p(\xi) \ge \mathcal{M}_p(\xi) \equiv \inf_\eta \sup_{\mathcal{F}_p(\xi)} \mathbb{E}\{[\eta(X+Z) - X]^2\},$$

the minimax risk over *all* estimators $\eta$. Further $\mathcal{M}_p(\xi) \ge \mathcal{M}_\infty(\xi)$, the minimax risk for estimation subject to the bounded mean constraint $|\mu| \le \xi$. That $\mathcal{M}_\infty(\xi) \to 1$ is shown, for example, in [DLM90, Eq. (2.6)]. $\qquad\square$

Of particular interest is the case of extremely sparse signals, which corresponds to the limit of small $\xi$. This regime was studied in detail in [DJ94] whose results we summarize below.

**Lemma 2.3** ([DJ94]). *As $\xi \to 0$ the minimax pair $(\nu_{\varepsilon_p(\xi), \mu_p(\xi)}, \tau_p(\xi))$ in Eq. (2.4) obeys*

$$
\begin{aligned}
\tau_p(\xi) &= \sqrt{2 \log(1/\xi^p)} \cdot \{1 + o(1)\}, \\
\mu_p(\xi) &= \sqrt{2 \log(1/\xi^p)} \cdot \{1 + o(1)\}, \\
\varepsilon_p(\xi) &= \left(\frac{\xi^2}{2 \log(1/\xi^p)}\right)^{p/2} \cdot \{1 + o(1)\}.
\end{aligned}
$$

*Further, the minimax mean square error is given, in the same limit, by*

$$M_p(\xi) = (2 \log(1/\xi^p))^{1-p/2} \xi^p \cdot \{1 + o(1)\}. \tag{2.11}$$

The asymptotics for $M_p(\xi)$ in the last lemma imply the following behavior of the inverse function as $m \to 0$:

$$M_p^{-1}(m) = \left(2 \log(1/m)\right)^{1/2 - 1/p} m^{1/p} \cdot \{1 + o(1)\}. \tag{2.12}$$

11

# 3 The asymptotic LASSO risk

In this section we discuss the high-dimensional limit of the LASSO mean square error for a given sequence of instances $\mathsf{S} = (I_{n,N})$. Our treatment is mainly a summary of results proved in [BM10] and [DMM10], adapted to the current context.

## 3.1 Convergent Sequences, and their AMSE

We introduced the notion of sequence of instances as a very general, almost structure-free notion; but certain special sequences play a distinguished role.

**Definition 3.1. Convergent sequence of problem instances.** *The sequence of problem instances* $\mathsf{S} = \{(x_0^{(N)}, z^{(n)}, A^{(n,N)})\}_{n,N}$ *is said to be a* convergent sequence *if* $n/N \to \delta \in (0, \infty)$, *and in addition the following conditions hold:*

(a) **Convergence of object marginals.** *The empirical distribution of the entries of* $x_0^{(N)}$ *converges weakly to a probability measure* $\nu$ *on* $\mathbb{R}$ *with bounded second moment. Further* $N^{-1}\|x_0^{(N)}\|_2^2 \to \mathbb{E}_\nu X^2$.

(b) **Convergence of noise marginals.** *The empirical distribution of the entries of* $z^{(n)}$ *converges weakly to a probability measure* $\omega$ *on* $\mathbb{R}$ *with bounded second moment. Further* $n^{-1}\|z_i^{(n)}\|_2^2 \to \mathbb{E}_\omega Z^2 \equiv \sigma^2$.

(c) **Normalization of Matrix Columns.** *If* $\{e_i\}_{1 \le i \le N}$, $e_i \in \mathbb{R}^N$ *denotes the standard basis, then* $\max_{i \in [N]} \|A^{(n,N)} e_i\|_2$, $\min_{i \in [N]} \|A^{(n,N)} e_i\|_2 \to 1$, *as* $N \to \infty$ *where* $[N] \equiv \{1, 2, \ldots, N\}$.

*We shall say that* $\mathsf{S}$ *is a* convergent sequence of problem instances, *and will write* $\mathsf{S} \in CS(\delta, \nu, \omega, \sigma)$ *to make explicit the limit objects.*

Next we need to introduce or recall some notations. The mean square error for scalar soft thresholding was already introduced in the previous Section, cf. Eq. (2.5), and denoted by $\mathsf{mse}(\sigma^2; \nu, \tau)$. The second is the following *state evolution map*

$$\Psi(m; \delta, \sigma, \nu, \tau) \equiv \mathsf{mse}\left(\sigma^2 + \frac{1}{\delta}m; \nu, \tau\right), \tag{3.1}$$

This is the mean square error for soft thresholding, when the noise variance is $\sigma^2 + m/\delta$. The addition of the last term reflects the increase of 'effective noise' in compressed sensing as compared to simple denoising, due to the undersampling. In order to have a shorthand for the latter, we define *noise plus interference* to be

$$\mathsf{npi}(m; \delta, \sigma) = \sigma^2 + \frac{m}{\delta}. \tag{3.2}$$

Whenever the arguments $\delta, \sigma, \nu, \tau$ will be clear from the context in the above functions, we will drop them and write, with an abuse of notation $\Psi(m)$ and $\mathsf{npi}(m)$.

Finally, we need to introduce the following *calibration relation*. Given $\tau \in \mathbb{R}_+$, let $m_*(\tau)$ to be the largest positive solution of the fixed point equation

$$m = \Psi(m; \delta, \sigma, \nu, \tau) \tag{3.3}$$

(of course $m_*$ depends on $\delta, \sigma, \nu$ as well but we'll drop this dependence unless necessary). Such a solution is finite for all $\tau > \tau_0$ for some $\tau_0 = \tau_0(\delta)$. The corresponding LASSO parameter is then given by

$$\lambda(\tau) \equiv \tau \sqrt{\mathsf{npi}_*} \left[ 1 - \frac{1}{\delta} \mathbb{P}\{ |X + \sqrt{\mathsf{npi}_*} Z| \geq \tau \sqrt{\mathsf{npi}_*} \} \right] . \tag{3.4}$$

with $\mathsf{npi}_* = \mathsf{npi}(m_*(\tau))$. As shown in [BM10], $\tau \mapsto \lambda(\tau)$ establishes a bijection between $\lambda \in (0, \infty)$ and $\tau \in (\tau_1, \infty)$ for some $\tau_1 = \tau_1(\delta) > \tau_0(\delta)$.

The basic high-dimensional limit result can be stated as follows.

**Theorem 3.1.** *Let $\mathsf{S} = \{I_{n,N}\} = \{(x_0^{(N)}, z^{(n)}, A^{(n,N)})\}_{n,N}$ be a convergent sequence of problem instances, $\mathsf{S} \in CS(\delta, \sigma, \nu, \omega)$, and assume also that the matrices $A^{(n,N)}$ are sampled from $\mathrm{GAUSS}(n, N)$. Denote by $\hat{x}_\lambda^{(N)}$ the LASSO estimator for instance $I_{n,N}$, $\lambda \geq 0$ and let $\psi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a locally-Lipschitz function with $|\psi(x_1, x_2)| \leq C(1 + x_1^2 + x_2^2)$ for all $x_1, x_2 \in \mathbb{R}$.*

*Then, almost surely*

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^N \psi(\hat{x}_{\lambda,i}, x_{0,i}) = \mathbb{E}\left\{ \psi(\eta(X + \sqrt{\mathsf{npi}_*} Z; \tau_* \sqrt{\mathsf{npi}_*}), X) \right\}, \tag{3.5}$$

*where $\mathsf{npi}_* \equiv \mathsf{npi}(m_*)$, $Z \sim \mathsf{N}(0, 1)$ is independent of $X \sim \nu$, $\tau_* = \tau_*(\lambda)$ is given by the calibration relation described above, and $m_*$ is the largest positive solution of the fixed point equation $m = \Psi(m, \delta, \sigma, \nu, \tau_*)$.*

## 3.2 Discussion and further properties

In the next pages we will repeatedly use the shorthand $\mathrm{HFP}(\Psi)$ to denote the largest positive solution of the fixed point equation $m = \Psi(m; \delta, \sigma, \nu, \tau)$, where we may suppress the secondary parameters $(\delta, \sigma, \nu, \tau)$ and simply write $\Psi(m)$. Formally

$$\mathrm{HFP}(\Psi) \equiv \sup\{ m \geq 0 : \Psi(m) \geq m \}. \tag{3.6}$$

In order to emphasize the role of parameters $\delta, \sigma, \nu, \tau$, we may also write $\mathrm{HFP}(\Psi(\,\cdot\,; \delta, \sigma, \nu, \tau))$. We recall some basic properties of the mapping $\Psi$.

**Lemma 3.1** ([DMM09, DMM10]). *For fixed $\delta, \sigma, \nu, \tau$, the mapping $m \mapsto \Psi(m)$ defined on $[0, \infty)$ is continuous, strictly increasing and concave. Further $\Psi(0) \geq 0$ with $\Psi(0) = 0$ if and only if $\sigma = 0$. Finally, there exists $\tau_0 = \tau_0(\delta)$ such that $\lim_{m \to \infty} \Psi'(m) < 1$ if and only if $\tau > \tau_0$.*

By specializing Theorem 3.1 to the case $\psi(x_1, x_2) = (x_1 - x_2)^2$ and using the fixed point condition $m_* = \Psi(m_*; \delta, \sigma, \nu, \tau_*)$ we obtain immediately the following.

**Corollary 3.1.** *Let $\mathsf{S} \in CS(\delta, \sigma, \nu, \omega)$ be a convergent sequence of problem instances, and further assume that $A^{(n,N)} \sim \mathrm{GAUSS}(n, N)$. Denote by $\hat{x}_\lambda^{(N)}$ the LASSO estimator for problem instance $I_{n,N}$, with $\lambda \geq 0$ . Then, almost surely*

$$\lim_{N \to \infty} \frac{1}{N} \|\hat{x}_\lambda - x_0\|_2^2 = m_*, \tag{3.7}$$

*where $m_* = \mathrm{HFP}(\Psi(\,\cdot\,; \delta, \sigma, \nu, \tau_*))$, and $\tau_* = \tau_*(\lambda)$ is fixed by the calibration relation (3.4).*

## 3.3 AMSE over General Sequences

Corollary 3.1 determines the asymptotic mean square error for convergent sequences $\mathsf{S} \in CS(\delta, \sigma, \nu)$. The resulting expression depends on $\delta, \sigma, \nu$, and is denoted $\mathrm{AMSE}_{SE}(\lambda; \delta, \sigma, \nu)$. We have

$$\mathrm{AMSE}_{SE}(\lambda; \delta, \sigma, \nu) = \mathrm{HFP}(\Psi(\,\cdot\,; \delta, \sigma, \nu, \tau_*)). \tag{3.8}$$

The introduction considered instead the asymptotic mean square error $\mathrm{AMSE}(\lambda; \mathsf{S})$ along general, not necessarily convergent sequences of problem instances in the standard $\ell_p$ problem suite $\mathsf{S} \in \mathcal{S}_p(\delta, \xi, \sigma)$, cf. Eq. (1.4). Given a sequence $\mathsf{S} \in \mathcal{S}_p(\delta, \xi, \sigma)$, we let

$$\mathrm{AMSE}(\lambda; \mathsf{S}) = \lim \sup_{N \to \infty} \frac{1}{N} \mathbb{E}\{\|\widehat{x}_\lambda^{(N)} - x_0^{(N)}\|^2\}. \tag{3.9}$$

Below we will often omit the subscript $SE$ on $AMSE_{SE}$, thereby using the same notation for the state evolution quantity (3.8) and the sequence quantity (3.9). This abuse is justified by the following key fact. The asymptotic mean square error along *any* sequence of instances can be represented by the formula $\mathrm{AMSE}_{SE}(\lambda; \delta, \nu, \sigma)$, for a suitable $\nu$ – provided the sensing matrices $A^{(n,N)}$ have i.i.d. Gaussian entries. Before stating this result formally, we recall that the definition of sparsity class $\mathcal{F}_p(\xi)$ was given in Eq. (2.3).

**Proposition 3.1.** *Let $\mathsf{S}$ be any* (not necessarily convergent) *sequence of problem instances in $\mathcal{S}_p(\delta, \xi, \sigma)$. Then there exists a probability distribution $\nu \in \mathcal{F}_p(\xi)$ such that*

$$\mathrm{AMSE}(\lambda; \mathsf{S}) = \mathrm{AMSE}_{SE}(\lambda; \delta, \nu, \sigma), \tag{3.10}$$

*and both sides are given by the fixed point of the one-dimensional map $\Psi$, namely* $\mathrm{HFP}(\Psi(\,\cdot\,; \delta, \sigma, \nu, \tau_*))$. *Further, for each $\varepsilon > 0$,*

$$\lim \sup_{N \to \infty} \mathbb{P}\Big\{\frac{1}{N}\|\widehat{x}_\lambda^{(N)} - x_0^{(N)}\|_2^2 \geq \mathrm{AMSE}_{SE}(\lambda; \delta, \nu, \sigma) + \varepsilon\Big\} = 0. \tag{3.11}$$

*Conversely, for any $\nu \in \mathcal{F}_p(\xi)$, there exists a sequence of instances $\mathsf{S} \in \mathcal{S}_p(\delta, \xi, \sigma)$, such that $\mathrm{AMSE}(\lambda; \mathsf{S}) = \mathrm{AMSE}(\lambda; \delta, \nu, \sigma)$ along that sequence.*

*Proof.* Given the sequence of problem instances, $\mathsf{S} = \{x_0^{(N)}, z^{(n)}, A^{(n,N)}\}_{n,N}$, extract a subsequence along which the expected mean square error has a limit equal to the $\lim \sup$ in Eq. (1.4). We will then extract a further subsequence that is a convergent subsequence of problem instances, in the sense of Definition 3.1, hence proving the direct part of our claim, by virtue of Corollary 3.1. (Convergence of the expectation of $\|\widehat{x}_\lambda^{(N)} - x_0^{(N)}\|^2/N$ follows from almost sure convergence together with the fact that $\|x_0^{(N)}\|^2/N$ is uniformly bounded by assumption and $\|\widehat{x}_\lambda^{(N)}\|^2/N$ is uniformly bounded by Lemma 3.3 in [BM10].)

Let $\nu_{x_0,N}$ be the empirical distribution of $x_0^{(N)}$ as in (2.2). Since $\mathsf{S} \in \mathcal{S}_p(\delta, \xi, \sigma)$, we have $\nu_{x_0,N}(|X|^p) \leq \xi^p$ hence the family $\{\nu_{x_0,N}\}$ is tight, and along a further subsequence the empirical distributions of $x_0^{(N)}$ converge weakly, to a limit $\nu$, say. Again by $\mathsf{S} \in \mathcal{S}_p(\delta, \xi, \sigma)$, the empirical distributions of $z^{(n)}$ are tight (assumption $\mathbf{z} \in \mathcal{Z}^2(\sigma)$ entails $\|z^{(n)}\|^2/n \to \sigma^2$); we extract yet another subsequence along which they converge, to $\omega$, say.

14

We are left with a subsequence we shall label $\{(n_k, N_k)\}_{k\geq 1}$. We wish to prove for this sequence (a)-(c) of Definition 3.1. Property (c) in Definition 3.1, the convergence of column norms, is well known to hold for random matrices with iid Gaussian entries (and easy to show). We are left to show (a) and (b), i.e. that $\nu_{x_0,N_k}(X^2) \to \nu(X^2)$ and $\nu_{z,n_k}(X^2) \to \omega(X^2)$ along this sequence. Convergence of the second moments follows since

$$\lim_{k\to\infty} \nu_{x_0,N_k}(X^2) = \lim_{k\to\infty} \nu_{x_0,N_k}(X^2 \mathbb{I}_{\{|X|\leq M\}}) + \lim_{k\to\infty} \nu_{x_0,N_k}(X^2 \mathbb{I}_{\{|X|> M\}})$$
$$= \nu(X^2 \mathbb{I}_{\{|X|\leq M\}}) + \mathsf{err}_M$$

where we used the dominated convergence theore, where, by the uniform integrability property of sequences $\mathbf{x}$ in $\mathcal{X}_p(\xi)$, $\mathsf{err}_M \leq \epsilon_M \downarrow 0$ as $M \to \infty$.

The limit in probability (3.11) follows by very similar arguments and we omit it here.

The converse is proved by taking $x_0^{(N)}$ to be a vector with iid components $x_0^{(N)} \sim \nu$. The empirical distributions $\nu_N$ then converge almost surely to $\nu$ by the Glivenko-Cantelli theorem. Convergence of second moments follows from the strong law of large numbers. $\square$

## 3.4 Intuition and relation to AMP algorithm

Theorem 3.1 implies that, in the high-dimensional limit, vector estimation through the LASSO can be effectively understood in terms of $N$ uncoupled scalar estimation problems, provided the noise is augmented by an undersampling-dependent increment. A natural question is whether one can construct, starting from the vector of measurements $y = (y_1, \ldots, y_n)$ (which are intrisicaly 'joint' measurements of $x_1, \ldots, x_N$), a collection of $N$ uncoupled measurements of $x_1, \ldots, x_N$.

A deeper intuition about this question and Theorem 3.1 can be developed by considering the approximate message passing (AMP) algorithm first introduced in [DMM09]. At one given problem instance (i.e. frozen choice of $(n, N)$) we omit the superscript $(N)$. The algorithm produces a sequence of estimates $\{\widehat{x}^0, \widehat{x}^1, \widehat{x}^2 \ldots\}$ in $\mathbb{R}^N$, by letting $\widehat{x}^0 = 0$ and, for each $t \geq 0$

$$z^t = y - A\widehat{x}^t + \frac{\|\widehat{x}^t\|_0}{n} z^{t-1} \tag{3.12}$$
$$\widehat{x}^{t+1} = \eta(\widehat{x}^t + A^T z^t; \theta_t), \tag{3.13}$$

where $\|\widehat{x}^t\|_0$ is the size of the support of $\widehat{x}^t$. Here $\{z^t\}_{t\geq 0} \subseteq \mathbb{R}^n$ is a sequence of residuals and $\theta_t$ a sequence of thresholds.

As shown in [BM11], the vector $\widehat{x}^t + A^T z^t$ is distributed asymptotically (large $t$) as $x_0 + w^t$ with $w^t \in \mathbb{R}^N$ a vector with i.i.d. components $w_i^t \sim \mathsf{N}(0, \sigma_t^2)$ independent of $x_0$. (Here the convergence is to be understood in the sense of finite-dimensional marginals.) In other words, *the vector $\widehat{x}^t + A^T z^t$ produced by the AMP algorithm is effectively a vector of i.i.d. uncoupled observations of the signal $x_0$.*

The second key point is that the AMP algorithm is tightly related to the LASSO. First of all, fixed points of AMP (for a fixed value of the threshold $\theta_t = \theta_*$) are minimizers of the LASSO cost function and viceversa, provided the $\theta_*$ is calibrated with the regularization parameter $\lambda$ according to the following relation

$$\lambda = \theta_* \cdot \left(1 - \frac{\|\widehat{x}_\lambda\|_0}{n}\right), \tag{3.14}$$

15

with $\hat{x}_\lambda$ the LASSO minimizer or –equivalently– the AMP fixed point. Finally, [BM10] proved that (for Gaussian sensing matrices $A$), the AMP estimates do converge to the LASSO minimizer provided the sequence of thresholds is chosen according to the policy

$$\theta_t = \tau \, \sigma_t \,, \tag{3.15}$$

for a suitable $\alpha > 0$ depending on $\lambda$ [BM10, DMM10]. Finally, the effective noise-plus-interference level $\sigma_t$ can be estimated in several ways, a simple one being $\hat{\sigma}_t^2 = \|z_t\|^2/n$.

# 4 Minimax MSE over $\ell_p$ Balls, Noiseless Case

In this section we state results for the noiseless case, $y = Ax_0$, where $A$ is $n \times N$ and $x_0$ obeys an $\ell_p$ constraint. As mentioned in the introduction, our results hold in the asymptotic regime where $n/N \to \delta \in (0,1)$.

## 4.1 Main Result

Let $\mathsf{S} \equiv \{(x_0^{(N)}, z^{(n)}, A^{(n,N)})\}_{n,N}$ be a sequence of *noiseless* problem instances ($z^{(n)} = 0$: no noise is added to the measurements) with Gaussian sensing matrices $A^{(n,N)} \sim \text{GAUSS}(n, N)$. Define the minimax LASSO mean square error as

$$M_p^*(\delta, \xi) \equiv \sup_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,0)} \inf_{\lambda \in \mathbb{R}_+} \text{AMSE}(\lambda; \mathsf{S}) \,. \tag{4.1}$$

**Theorem 4.1.** *Fix $\delta \in (0,1)$, $\xi > 0$. The minimax AMSE obeys:*

$$M_p^*(\delta, \xi) = \frac{\delta \xi^2}{M_p^{-1}(\delta)^2}. \tag{4.2}$$

*Further we have:*
**Minimax Threshold.** *The minimax threshold $\lambda^*(\delta, \xi)$ is given by the calibration relation (3.4) with $\tau = \tau^*(\delta, \xi)$ determined as follows (notice in particular that this is independent of $\xi$):*

$$\tau^*(\delta, \xi) = \tau_p(M_p^{-1}(\delta)) \,. \tag{4.3}$$

**Least Favorable $\nu$.** *The least-favorable distribution is a 3-point distribution $\nu^* = \nu_{p,\delta,\xi}^* = \nu_{\varepsilon^*, \mu^*}$ (cf. Eq. (2.6)) with*

$$\mu^*(\delta, \xi) = \frac{\xi}{M_p^{-1}(\delta)} \, \mu_p(M_p^{-1}(\delta)) \,, \quad \varepsilon^*(\delta, \xi) = \frac{\xi^p}{(\mu^*)^p} \,. \tag{4.4}$$

**Saddlepoint.** *The above quantities obey a saddlepoint relation. Put for short* $\text{AMSE}(\lambda; \nu)$ *in place of* $\text{AMSE}(\lambda; \delta, \nu, 0)$, *The minimax AMSE obeys*

$$M_p(\delta; \xi) = \text{AMSE}(\lambda^*; \nu^*)$$

*and*

$$
\begin{aligned}
\text{AMSE}(\lambda^*; \nu^*) &\leq \text{AMSE}(\lambda; \nu^*), && \forall \lambda > 0 & (4.5) \\
&\geq \text{AMSE}(\lambda^*; \nu), && \forall \nu \in \mathcal{F}_p(\xi). & (4.6)
\end{aligned}
$$
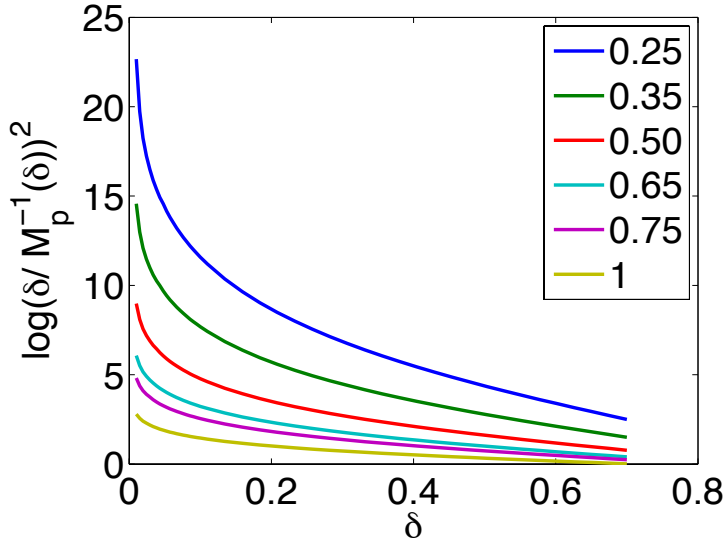
Figure 5: Minimax MSE $M_p^*(\delta, 1)$. We assume here $\xi = 1$; curves show log MSE as a function of $\delta$. Consistent with $\delta \to 0$ asymptotic theory, the curves are nearly scaled copies of each other.

## 4.2   Interpretation

Figure 5 presents the function $M_p^*(\delta, \xi = 1)$ on a logarithmic scale. As the reader can see, there is a substantial increase in the minimax risk as $\delta \to 0$, which agrees with our intuitive picture that the reconstruction becomes less accurate for small $\delta$ (high undersampling).

The asymptotic properties of $M_p^*(\delta, 1)$ in the high undersampling regime ($\delta \to 0$) can be derived using Lemma 2.3. From Eq. (2.12) we have

$$M_p^*(\delta, 1) = \delta^{1-2/p}(2\log(\delta^{-1}))^{2/p-1}\{1 + o_\delta(1)\}, \qquad \delta \to 0.$$

Hence, when plotting $\log M_p^*(\delta, 1)$, as we do here, we should see graphs of the form

$$\log M_p^*(\delta, 1) = (1 - 2/p) \cdot \left[\log(\delta^{-1}) - \log(\log(\delta^{-1})) - \log(2)\right] + o_\delta(1), \qquad \delta \to 0.$$

In particular the curves should look 'all the same' at small $\delta$, except for scaling; this is qualitatively consistent with Fig 5, even at larger $\delta$.

Another useful prediction can be obtained by working out the asymptotics of the minimax threshold $\lambda^*(\delta, \xi)$. Using Eq. (4.3) as well as the calibration relation (3.4), we get, as $\delta \to 0$,

$$\lambda^*(\delta, \xi) = \xi \cdot \left(\frac{2\log(1/\delta)}{\delta}\right)^{1/p}\{1 + o_\delta(1)\}. \tag{4.7}$$

## 4.3 Proof of Theorem 4.1

We will focus on proving Eq. (4.2), since the other points follow straightforwardly. By Proposition 3.1, we have the equivalent characterization

$$M_p^*(\delta, \xi) = \sup_{\nu \in \mathcal{F}_p(\xi)} \inf_{\lambda \in \mathbb{R}_+} \text{AMSE}(\lambda; \delta, \nu, \sigma = 0). \tag{4.8}$$

Further, by Corollary 3.1, we can use the mean square error expression given there, and because of the monotone nature of the calibration relation, we can minimize over the threshold $\tau$ instead of $\lambda$. We get therefore

$$M_p^*(\delta, \xi) = \sup_{\nu \in \mathcal{F}_p(\xi)} \inf_{\tau \in \mathbb{R}_+} \text{AMSE}_{\text{SE}}(\tau; \delta, \nu, 0), \tag{4.9}$$

where

$$\text{AMSE}_{\text{SE}}(\tau; \delta, \nu, 0) = m,$$
$$m = \text{mse}(m/\delta; \nu, \tau). \tag{4.10}$$

Recall that $\mathcal{P}(\mathbb{R})$ denotes the class of all probability distribution functions on $\mathbb{R}$. Define the scaling operator $S_a : \mathcal{P}(\mathbb{R}) \to \mathcal{P}(\mathbb{R})$ by $(S_a \nu)(B) = \nu(B/a)$ for any Borel set $B$. For the family of operators $\{S_a : a > 0\}$ we have the group properties

$$S_a \cdot S_b = S_{a \cdot b}, \qquad S_1 = \mathbf{I}, \qquad S_a S_{a^{-1}} = S_1. \tag{4.11}$$

In particular by the last property, for any $a > 0$, the operator $S_a : \mathcal{P}(\mathbb{R}) \mapsto \mathcal{P}(\mathbb{R})$ is one-to-one.

With this notation, we have the scale covariance property of the soft-thresholding mean square error

$$\text{mse}(\sigma^2; \nu, \tau) = \sigma^2 \cdot \text{mse}(1; S_{1/\sigma}\nu, \tau), \tag{4.12}$$

transforming a general-noise-level problem into a noise-level-one problem. As a consequence of Lemma 3.1, the map $\sigma^2 \mapsto \text{mse}(\sigma^2; \nu, \tau)$ is (for fixed $\nu, \tau$) increasing and concave. Therefore, the map $\sigma^2 \mapsto \text{mse}(1; S_{1/\sigma}\nu, \tau)$ is strictly monotone decreasing. Also, the fixed point Eq. (4.10) can be rewritten as

$$\delta = \text{mse}(1; S_{\sqrt{\delta/m}}\nu, \tau), \tag{4.13}$$

where the solution is unique by strict monotonicity of $m \mapsto \text{mse}(1; S_{\sqrt{\delta/m}}\nu, \tau)$.

We will prove Eq. (4.2) by obtaining an upper and a lower bound for $M_p(\delta, \xi)$. In the following we assume without loss of generality that the infimum in Eq. (4.8) is achieved

$$M_p(\delta, \xi) = \text{AMSE}_{\text{SE}}(\tau_*; \nu_*, 0). \tag{4.14}$$

Further we will use the minimax conditions for soft thresholding, see Lemma 2.1:

$$\inf_{\tau \in \mathbb{R}_+} \text{mse}(1; \nu, \tau) \leq M_p(\xi), \qquad \forall \nu \in \mathcal{F}_p(\xi), \tag{4.15}$$

$$\sup_{\nu \in \mathcal{F}_p(\xi)} \text{mse}(1; \nu, \tau) \geq M_p(\xi), \qquad \forall \tau \in \mathbb{R}_+. \tag{4.16}$$
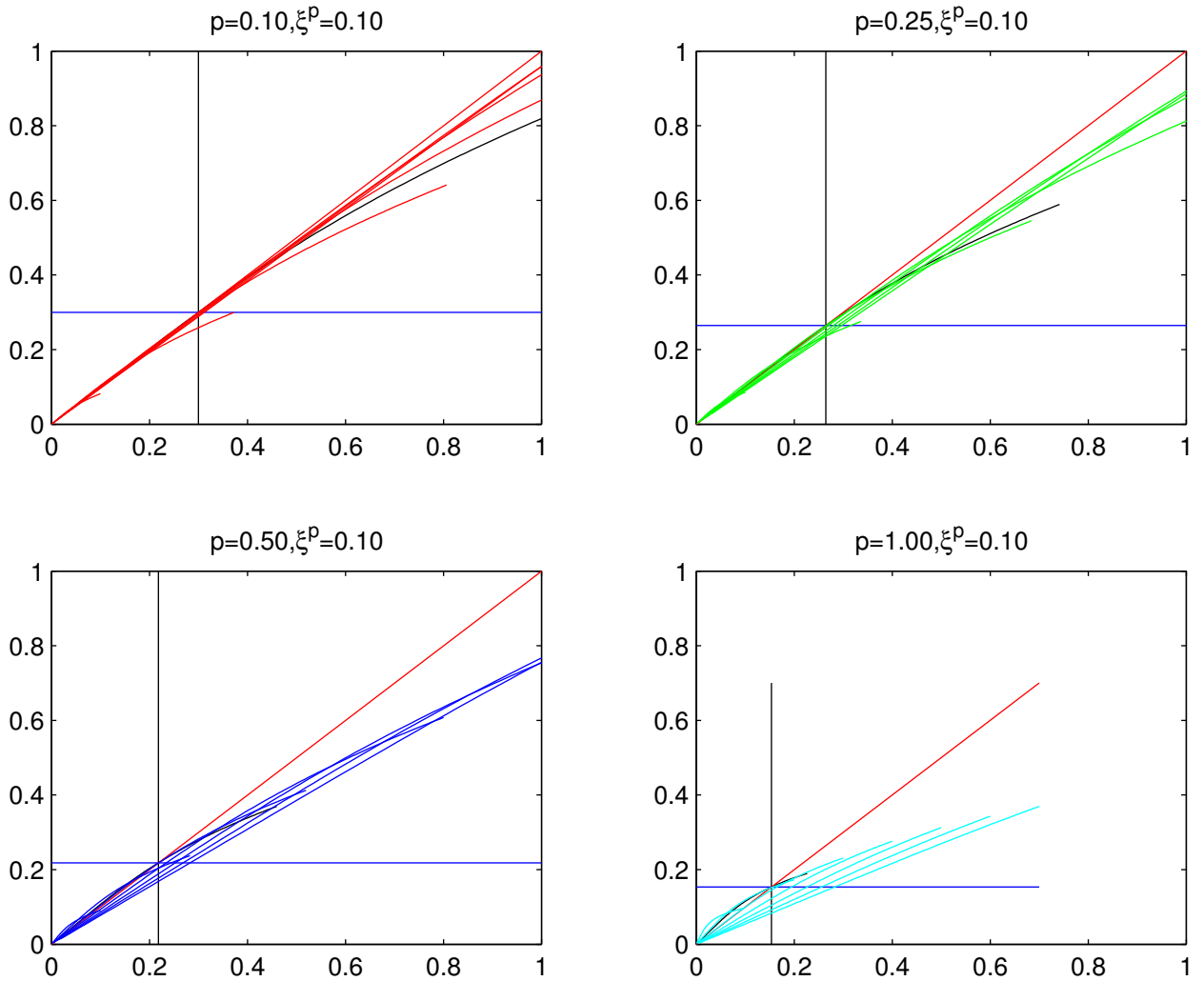
Figure 6: Illustration of the minimax fixed point property. Horizontal input MSE $m$. Vertical: output MSE $\Psi(m)$ for the state evolution map defined as per Eq. (3.1). Red diagonal: $\Psi(m) = m$. Black vertical Line: minimax HFP $M_p(\xi)$. Blue horizontal line: minimax MSE $M_p(m)$. Black curve: MSE map at minimax threshold value and least-favorable distribution. It crosses the diagonal at the minimax fixed point. Colored Curves. MSE maps at minimax threshold value and other three-point distributions. All other fixed points occur below $M_p(\xi)$.
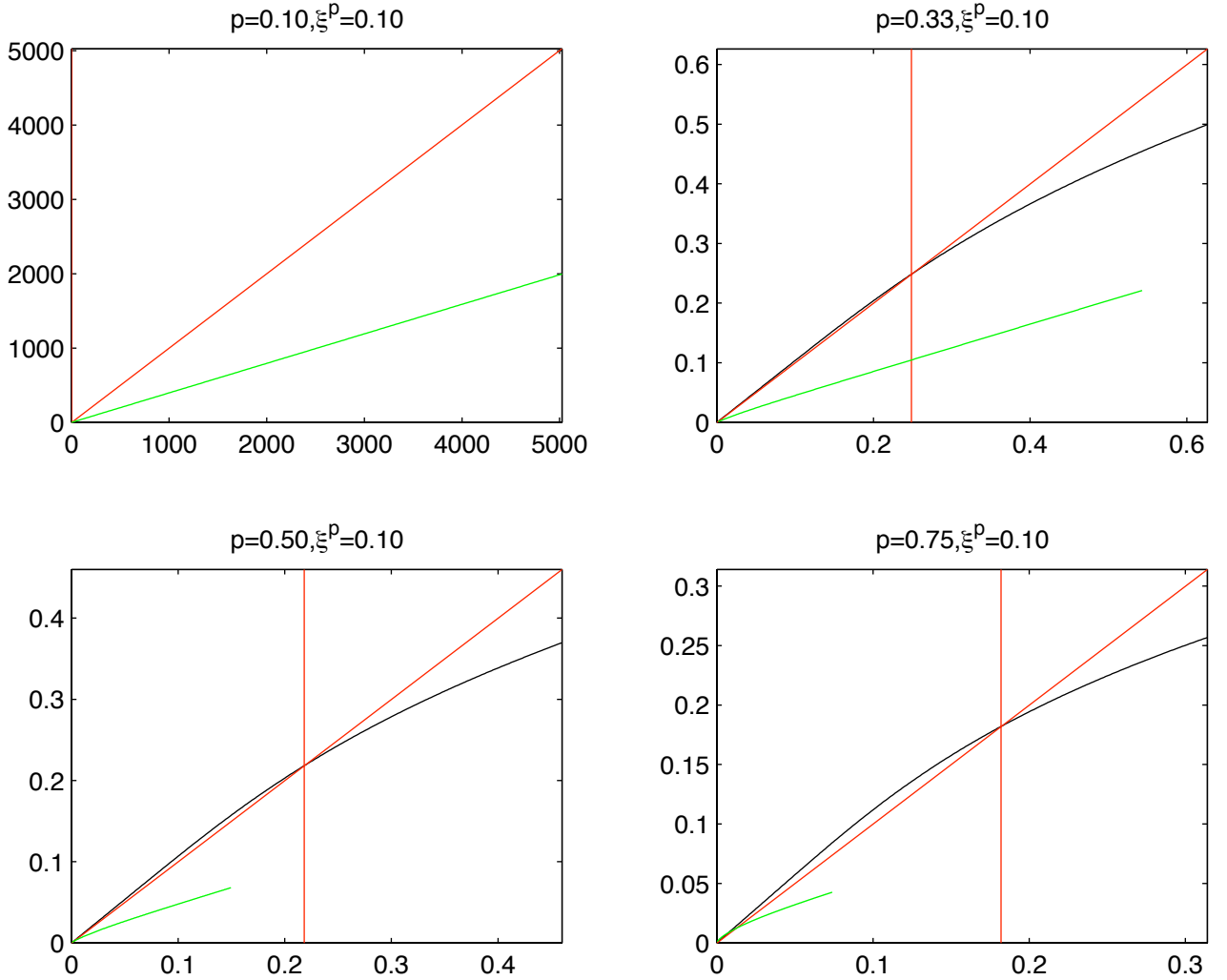
19

Figure 7: Comparisons of highest fixed point at power law distribution with minimax HFP. Horizontal input MSE $m$. Vertical: output MSE $\Psi(m)$ for the state evolution map defined as per Eq. (3.1). Red diagonal: $\Psi(m) = m$. Red vertical: Minimax MSE. Black curve: MSE map at minimax threshold value and least-favorable distribution. It crosses the diagonal at the minimax fixed point. Green Curve: MSE map with same threshold, taken at power law distribution calibrated to same $\mathbb{E}|X|^p = \xi^p$ constraint.

**Upper bound on $M_p(\delta, \xi)$.** Let $m_* = M_p(\delta, \xi) = \mathrm{AMSE}_{\mathrm{SE}}(\tau_*; \nu_*, 0)$. By Eq. (4.10) and (4.13) we have

$$\delta = \mathsf{mse}(1; S_{\sqrt{\delta/m_*}} \nu_*, \tau_*) = \inf_{\tau \in \mathbb{R}_+} \mathsf{mse}(1; S_{\sqrt{\delta/m_*}} \nu_*, \tau) . \tag{4.17}$$

The second equality follows because otherwise by there would exist $\tau_{**}$ with $\mathsf{mse}(1; S_{\sqrt{\delta/m}} \nu_*, \tau_{**})$ whence, by the monotonicity of $m \mapsto \mathsf{mse}(1; S_{\sqrt{\delta/m}} \nu_*, \tau_{**})$ it would follow that $\mathrm{AMSE}_{\mathrm{SE}}(\tau_{**}; \nu_*, 0) < \mathrm{AMSE}_{\mathrm{SE}}(\tau_*; \nu_*, 0)$ which violates the minimax property (4.14).

Next notice that $S_{\sqrt{\delta/m_*}} \nu_* \in \mathcal{F}_p(\sqrt{\delta/m_*}\,\xi)$ whence by Eq. (4.15), we get $\delta \leq M_p(\sqrt{\delta/m_*}\,\xi)$. By the monotonicity of $\xi \mapsto M_p(\xi)$ this yields

$$m_* \leq \frac{\delta\,\xi^2}{M_p^{-1}(\delta)^2} . \tag{4.18}$$

**Lower bound on $M_p(\delta, \xi)$.** Again by Eq. (4.10) and (4.13) we have

$$\delta = \mathsf{mse}(1; S_{\sqrt{\delta/m_*}} \nu_*, \tau_*) = \sup_{\nu \in \mathcal{F}_p(\xi)} \mathsf{mse}(1; S_{\sqrt{\delta/m_*}} \nu, \tau_*(\nu)) , \tag{4.19}$$

with $\tau_*(\nu)$ the optimal threshold for distribution $\nu$ and the second equality following by an argument similar to the one above (i.e. if this weren't true, there would be a *different* worst distribution $\nu_{**}$, reaching contradiction). But $\nu \in \mathcal{F}_p(\xi)$ implies $S_{\sqrt{\delta/m_*}} \nu \in \mathcal{F}_p(\sqrt{\delta/m_*}\,\xi)$, whence

$$\delta = \sup_{\nu \in \mathcal{F}_p(\xi\sqrt{\delta/m_*})} \mathsf{mse}(1; \nu, \tau_*(\nu)) \geq M_p(\sqrt{\delta/m_*}\,\xi) , \tag{4.20}$$

where the second inequality follows by Eq. (4.16). The proof is finished by using again the monotonicity of $\xi \mapsto M_p(\xi)$. $\qquad\square$

# 5 Minimax MSE over $\ell_p$ Balls, Noisy Case

In this section we generalize the results of the previous section to the case of noisy measurements with noise variance per coordinate equal to $\sigma^2$.

## 5.1 Main Result

Now let $\sigma > 0$ and consider sequences $S$ of *noisy* problem instances from the standard $\ell_p$ problem suite $S \in \mathcal{S}_p(\delta, \xi, \sigma)$; hence, in addition to the $\ell_p$ constraint $\|x_0^{(N)}\|_p^p \leq N\xi^p$ and each $A^{(n,N)} \sim \mathrm{GAUSS}(n, N)$, now the noise vectors $z^{(n)} \in \mathbb{R}^n$ are non-vanishing and have norms satisfying $\|z^{(n)}\|^2/n \to \sigma^2 > 0$.

We define the minimax LASSO asymptotic mean square error as

$$M_p^*(\delta, \xi, \sigma) \equiv \sup_{S \in \mathcal{S}_p(\delta, \xi, \sigma)} \inf_{\lambda \in \mathbb{R}_+} \mathrm{AMSE}(\lambda; S) . \tag{5.1}$$

By simple scaling of the problem we have, for any $\sigma > 0$,

$$M_p^*(\delta, \xi, \sigma) = \sigma^2\, M_p^*(\delta, \xi/\sigma, 1) , \tag{5.2}$$

an observation which will be used repeatedly in the following.

**Theorem 5.1.** *For any $\delta, \xi > 0$, let $m^* = m_p^*(\delta, \xi)$ be the unique positive solution of*

$$\frac{m^*}{1 + m^*/\delta} = M_p \left( \frac{\xi}{(1 + m^*/\delta)^{1/2}} \right) . \tag{5.3}$$

*Then the LASSO minimax mean square error $M_p^*$ is given by:*

$$M_p^*(\delta, \xi, \sigma) = \sigma^2 \cdot m^*(\delta, \xi/\sigma) . \tag{5.4}$$

*Further, denoting by $\xi^* \equiv (1 + m^*/\delta)^{-1/2}\xi/\sigma$, we have:*
**Least Favorable $\nu$.** *The least-favorable distribution is a 3-point mixture $\nu^* = \nu_{p,\delta,\xi,\sigma}^* = \nu_{\varepsilon^*, \mu^*}$ (cf. Eq. (2.6)) with*

$$\mu^*(\delta, \xi, \sigma) = \sigma \cdot (1 + m^*/\delta)^{1/2}\mu_p(\xi^*), \qquad \varepsilon^*(\delta, \xi, \sigma) = \frac{\xi^p}{(\mu^*)^p} , \tag{5.5}$$

*with $m^* = m^*(\delta, \xi/\sigma)$ given by the solution of Eq. (5.3)*
**Minimax Threshold.** *The minimax threshold $\lambda^*(\delta, \xi, \sigma)$ is given by the calibration relation (3.4) with $\tau = \tau^*(\delta, \xi, \sigma)$ determined as follows:*

$$\tau^*(\delta, \xi, \sigma) = \tau_p(\xi^*). \tag{5.6}$$

*with $\tau_p(\cdot)$ the soft thresholding minimax threshold, $\xi^* \equiv (1 + m^*/\delta)^{-1/2}\xi/\sigma$ and $\nu = \nu^*$ is the least favorable distribution given above.*
**Saddlepoint.** *The above quantities obey a saddlepoint relation. Put for short $\mathrm{AMSE}(\lambda; \nu)$ in place of $\mathrm{AMSE}(\lambda; \delta, \nu, \sigma)$. The minimax AMSE obeys*

$$M_p(\delta, \xi, \sigma) = \mathrm{AMSE}(\lambda^*; \nu^*) ,$$

*and*

$$
\begin{aligned}
\mathrm{AMSE}(\lambda^*; \nu^*) &\leq \mathrm{AMSE}(\lambda; \nu^*), && \forall \lambda > 0 && (5.7) \\
&\geq \mathrm{AMSE}(\lambda^*; \nu), && \forall \nu \in \mathcal{F}_p(\xi). && (5.8)
\end{aligned}
$$

## 5.2   Interpretation

Figure 8 provides a concrete illustration of Theorem 5.1. For various sparsity levels $\xi$ and undersampling factors $\delta$, the mean square error $M_p(\delta, \xi, \sigma)$ can be easily computed. As expected, the result is monotone increasing in $\xi$ and decreasing in $\delta$. For a given target mean square error, such plots allow to determine the required number of linear measurements.

Equations (5.3) and (5.4) are somewhat more complex that their noiseless counterpart. For this reason, it is instructive to work out the $\sigma \to 0$ limit $M_p^*(\delta, \xi, \sigma)$. By the basic scaling relation (5.4), this is equivalent to computing the $\xi \to \infty$ limit of $M_p^*(1, \delta, \xi) = m^*(\delta, \xi)$. Considering Eq. (5.3), it is easy to show that, for large $\xi$

$$m^*(\delta, \xi) = c_0(\delta)\xi^2 + c_1(\delta) + O(\xi^{-2}) \equiv c(\delta, \xi)\xi^2 .$$
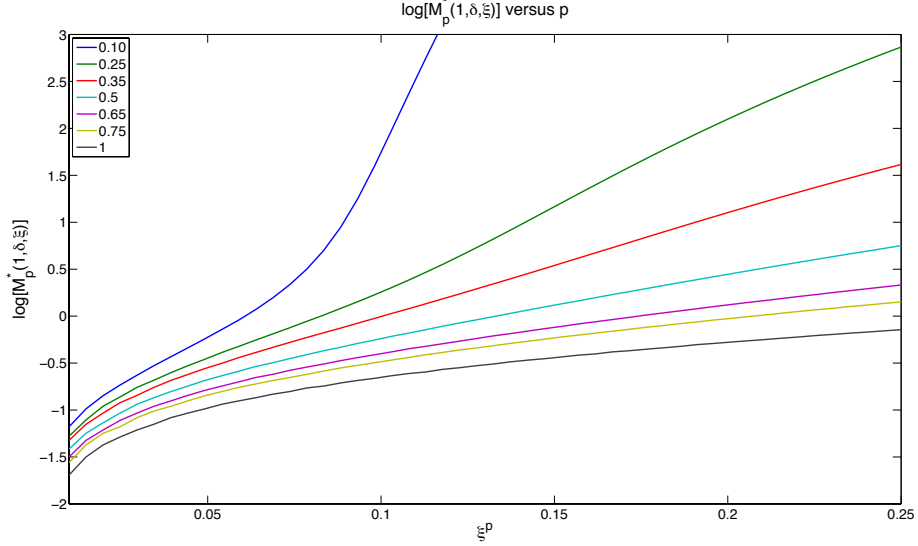
Figure 8: Minimax MSE $M_p^*(\delta, \xi, 1)$, noisy case $\sigma = 1$. We assume here $\delta = 1/4$; curves show MSE as a function of $\xi$.

Substituting in Eq. (5.3)

$$\delta \left(1 + \frac{\delta}{c\xi^2}\right)^{-1} = M_p \left((\delta/c)^{1/2} \left(1 + \frac{\delta}{c\xi^2}\right)^{-1/2}\right),$$

whence expanding for large $\xi$

$$\delta - \frac{\delta^2}{c_0 \xi^2} + O(\xi^{-4}) = M_p((\delta/c_0)^{1/2}) - \frac{\delta^{1/2}}{2\xi^2 c_0^{3/2}} M_p'((\delta/c_0)^{1/2}) (c_1 + \delta) + O(\xi^{-4}).$$

Imposing each order to vanish we get

$$c_0(\delta) = \frac{\delta}{M_p^{-1}(\delta)}, \tag{5.9}$$

$$c_1(\delta) = \frac{2\sqrt{c_0 \delta}}{M_p'((\delta/c_0)^{1/2})} - \delta. \tag{5.10}$$

Our calculations can be summarized as follows.

**Corollary 5.1.** *Fix a radius parameter $\xi$. As $\sigma^2 \to 0$, the asymptotic LASSO minimax mean square error behaves as*

$$M_p^*(\delta, \xi, \sigma) = \xi^2 c_0(\delta) + \sigma^2 c_1(\delta) + O(\sigma^4/\xi^2), \tag{5.11}$$

*with $c_0$ and $c_1$ determined by Eqs. (5.9) and (5.10). In particular, in the high undersampling regime $\delta \to 0$, we get*

$$c_1(\delta) = \frac{2}{p}\{1 + o_1(\delta)\}. \tag{5.12}$$

The derivation of the asymptotic behavior (5.12) is a straightforward calculus exercise, using Lemma 2.3.

The last Corollary shows that the noiseless case, cf. Theorem 4.1 and Eq. (4.2), is recovered as a special case of the noisy case treated in this section. Further leading corrections due to small noise $\sigma^2 \ll \xi^2$ are explicitly described by the coefficient $c_1(\delta)$ given in Eq. (5.10).

An alternative asymptotic of interest consists in fixing the noise level $\sigma$, and letting $\xi/\sigma \to 0$. In this regime the solution of Eq. (5.3) yields, using Lemma 2.3,

$$m^*(\delta, \xi) = (2\log(1/\xi^p))^{1-p/2}\xi^p \cdot \{1 + o(1)\} \,. \tag{5.13}$$

Substituting this expression in Theorem 5.1, we obtain the following.

**Corollary 5.2.** *Fix a noise parameter $\sigma^2 > 0$. As $\xi \to 0$, the asymptotic LASSO minimax mean square error behaves as*

$$M_p^*(\delta, \xi, \sigma) = \sigma^{2-p}\xi^p \cdot \left\{2\log\left((\sigma/\xi)^p\right)\right\}^{1-p/2} \cdot \{1 + o(1)\} \,, \tag{5.14}$$

*Further the minimax threshold value is given, in this limit, by*

$$\lambda^* = \sigma \cdot \sqrt{2\log\left((\sigma/\xi)^p\right)} \left\{1 + o(1)\right\} \,. \tag{5.15}$$

## 5.3 Proof of Theorem 5.1

The argument is structurally similar to the noiseless case. We will focus again on proving the asymptotic expression for minimax error given in Eq. (5.4), since the other points of the theorem follow easily. Using Proposition 3.1 and Corollary 3.1, the asymptotic mean square error can be replaced by the expression given there and the minimization over $\lambda$ can be replaced by a minimization over $\tau$:

$$M_p^*(\delta, \xi, \sigma) = \sup_{\nu \in \mathcal{F}_p(\xi)} \inf_{\tau \in \mathbb{R}_+} \mathrm{AMSE}_{\mathrm{SE}}(\tau; \delta, \nu, \sigma) \,, \tag{5.16}$$

where

$$\mathrm{AMSE}_{\mathrm{SE}}(\tau; \delta, \nu, \sigma) = m \,,$$
$$m = \mathsf{mse}\left(\sigma^2 + m/\delta; \nu, \tau\right) \,. \tag{5.17}$$

By virtue of the scaling relation (5.2), we can focus on the case $\sigma^2 = 1$. Define, for all $m < \delta$

$$\mathsf{n}(m) \equiv (1 + m/\delta)^{-1/2} \,. \tag{5.18}$$

We then have, applying Eq. (5.17) for the case $\sigma = 1$,

$$\frac{m}{1 + m/\delta} = \mathsf{mse}\left(1; S_{\mathsf{n}(m)}\nu, \tau\right) \,. \tag{5.19}$$

Notice that $m \mapsto m/(1 + m/\delta)$ is monotone increasing, and $m \mapsto \mathsf{mse}\left(1; S_{\mathsf{n}(m)}\nu, \tau\right)$ is monotone decreasing (because $a^2 \mapsto \mathsf{mse}\left(1; S_{1/a}\nu, \tau\right)$ is decreasing as mentioned in the

previous section). Hence this equation has a unique non-negative solution provided $\delta > \mathsf{mse}(1; \delta, \tau)$, which happens for all $\tau > \tau_0(\delta)$.

Assume without loss of generality that the minimax risk is achieved by the pair $(\tau_*, \nu_*)$. Then

$$M_p^*(\delta, \xi, 1) = \mathrm{AMSE}_{\mathrm{SE}}(\tau_*; \delta, \nu_*, 1) = m_* . \tag{5.20}$$

Then $m_*$ satisfies Eq. (5.19) with $\tau = \tau_*$ and $\nu = \nu_*$.

**Upper bound on $M_p(\delta, \xi, 1)$.** By the last remarks, we have

$$\frac{m_*}{1 + m_*/\delta} = \mathsf{mse}\big(1; S_{\mathsf{n}(m_*)}\nu_*, \tau_*\big) = \inf_{\tau \in \mathbb{R}_+} \mathsf{mse}\big(1; S_{\mathsf{n}(m_*)}\nu_*, \tau\big) . \tag{5.21}$$

The second equality follows from Eq. (5.20). Indeed if the equality did not hold, we could find $\tau_{**} \in \mathbb{R}_+$ such that $\mathsf{mse}\big(1; S_{\mathsf{n}(m_*)}\nu_*, \tau_{**}\big) < \mathsf{mse}\big(1; S_{\mathsf{n}(m_*)}\nu_*, \tau_*\big)$. But by the monotonicity of $m \mapsto m/(1 + m/\delta)$ and of $m \mapsto \mathsf{mse}\big(1; S_{\mathsf{n}(m)}\nu_*, \tau_{**}\big)$, this would mean that the corresponding fixed point $m_{**}$ is strictly smaller than $m_*$. This would contradict the minimax assumption.

Since $S_{\mathsf{n}(m_*)}\nu_* \in \mathcal{F}_p(\mathsf{n}(m_*)\xi)$ we can now apply Eq. (4.15), getting

$$\frac{m_*}{1 + m_*/\delta} \leq M_p(\mathsf{n}(m_*)\xi) . \tag{5.22}$$

Again by monotonicity of $m \mapsto m/(1 + m/\delta)$ and of $\xi \mapsto M_p(\xi)$, this means that $m_*$ is upper bounded by the solution of Eq. (5.3).

**Lower bound on $M_p(1, \delta, \xi)$.** Applying again Eq. (5.19) and an analogous argument as above, we have

$$\frac{m_*}{1 + m_*/\delta} = \sup_{\nu \in \mathcal{F}_p(\xi)} \mathsf{mse}\big(1; S_{\mathsf{n}(m_*)}\nu, \tau_*(S_{\mathsf{n}(m_*)}\nu)\big) . \tag{5.23}$$

In the last expression $\tau_*(S_{\mathsf{n}(m_*)}\nu)$ is the optimal (minimal MSE) threshold for distribution $S_{\mathsf{n}(m_*)}\nu$. For $\nu \in \mathcal{F}_p(\xi)$, $S_{\mathsf{n}(m_*)}\nu \in \mathcal{F}_p(\mathsf{n}(m_*)\xi)$. Further the map $S_{\mathsf{n}(m_*)} : \mathcal{F}_p(\xi) \to \mathcal{F}_p(\mathsf{n}(m_*)\xi)$ is bijective. We thus have

$$\frac{m_*}{1 + m_*/\delta} = \sup_{\nu \in \mathcal{F}_p(\mathsf{n}(m_*)\xi)} \mathsf{mse}\big(1; \nu, \tau_*(\nu)\big) . \tag{5.24}$$

By Eq. (4.16), we thus have

$$\frac{m_*}{1 + m_*/\delta} \leq M_p(\mathsf{n}(m_*)\xi) , \tag{5.25}$$

which implies that $m_*$ is upper bounded by the solution of Eq. (5.3). This finishes our proof. □

# 6 Weak $p$-th Moment Constraints

Our results for $\ell_p$ constraints have natural counterparts for weak $\ell_p$ constraints. We recall a standard definition for the weak-$\ell_p$ quasi-norm $\|x\|_{w\ell_p}$. For a vector $x \in \mathbb{R}^N$, let $T_x(t) \equiv \{i \in \{1, \ldots, N\} : |x_i| \geq t\}$ index the entries of $x$ with amplitude above threshold $t$. Denoting by $|S|$ the cardinality of set $S$, we define

$$\|x\|_{w\ell_p} \equiv \max_{t \geq 0} \left[ t|T_x(t)|^{1/p} \right], \tag{6.1}$$

By Markov's inequality $\|x\|_{w\ell_p} \leq \|x\|_p$: the weak $\ell_p$ quasi-norm is indeed weaker than the $\ell_p$ norm (quasi norm, if $p < 1$). Weak-$\ell_p$ norms arise frequently in applied harmonic analysis, as we discuss below.

As the reader no doubt expects, we can define a weak $\ell_p$ analogue to the $\ell_p$ case.

**Definition 6.1.** • **Weak $\ell_p$ constraint.** *A sequence* $\mathbf{x}_0 = (x_0^{(N)})$ *belongs to* $\mathcal{X}_p^w(\xi)$ *if* (i) $\|x_0^{(N)}\|_{w\ell_p}^p \leq N\xi^p$, *for all $N$; and* (ii) *there exists a sequence* $B = \{B_M\}_{M \geq 0}$ *such that* $B_M \to 0$, *and for every $N$,* $\sum_{i=1}^N (x_{0,i}^{(N)})^2 \mathbb{I}(|x_{0,i}^{(N)}| \geq M) \leq B_M N$.

• **Standard Weak-$\ell_p$ Problem Suite.** *Let* $\mathcal{S}_p^w(\delta, \xi, \sigma)$ *denote the class of sequences of problem instances* $I_{n,N} = (x_0^{(N)}, z^{(n)}, A^{(n,N)})$ *built from objects in weak $\ell_p$; in detail:* (i) $n/N \to \delta$; (ii) $\mathbf{x}_0 \in \mathcal{X}_p^w(\xi)$; (iii) $\mathbf{z} \in \mathcal{Z}_2(\sigma)$, *and* (iv) $A^{(n,N)} \in \textsc{Gauss}(n, N)$.

## 6.1 Scalar Minimax Thresholding under Weak $p$-th Moment Constraints

The class of probability distributions corresponding to instances in the weak-$\ell_p$ problem suite is

$$\mathcal{F}_p^w(\xi) \equiv \left\{ \nu \in \mathcal{P}(\mathbb{R}) : \sup_{t \geq 0} t^p \cdot \nu(\{|X| \geq t\}) \leq \xi^p \right\}. \tag{6.2}$$

In particular, given a sequence $\mathbf{x}_0 \in \mathcal{X}_p^w(\xi)$, the empirical distribution of each $x_0^{(N)}$ is in $\mathcal{F}_p^w(\xi)$.

As in section 2, we denote by $\mathsf{mse}(\sigma^2; \nu, \tau)$ the mean square error of scalar soft thresholding for a given signal distribution $\nu$.

**Definition 6.2.** *The* **minimax mean squared error** *under the weak $p$-th moment constraint is*

$$M_p^w(\xi) = \inf_{\tau \in \mathbb{R}_+} \sup_{\nu \in \mathcal{F}_p^w(\xi)} \mathbb{E}\left\{ \left[ \eta(X + Z; \tau) - X \right]^2 \right\}, \tag{6.3}$$

*where the expectation on the right hand side is taken with respect to $X \sim \nu$ and $Z \sim \mathsf{N}(0, 1)$, $X$ and $Z$ independent.*

The collection of probability measures $\mathcal{F}_p^w(\xi)$ has a distinguished element – a most dispersed one. In fact define the envelope function

$$H_{p,\xi}(t) = \inf_{\nu \in \mathcal{F}_p^w(\xi)} \nu(\{|X| \le t\}); \tag{6.4}$$

the envelope of achievable dispersion of the probability mass for elements of $\mathcal{F}_p^w(\xi)$. This envelope can be computed explicitly, yielding

$$H_{p,\xi}(t) = \begin{cases} 0 & \text{for } t < \xi, \\ 1 - (\xi/t)^p & \text{for } t \ge \xi. \end{cases} \tag{6.5}$$

Indeed it is clear by definition that $\nu(\{|X| \le t\}) \ge H_{p,\xi}(t)$. Further defining the CDF

$$F_{p,\xi}^w(x) = \begin{cases} \frac{1}{2} + \frac{1}{2}H_p(|x|) & \text{for } x \ge 0, \\ \frac{1}{2}H_p(|x|) & \text{for } x < 0. \end{cases} \tag{6.6}$$

and letting $\nu_{p,\xi}$ be the corresponding measure, we get for any $t \ge 0$, $\nu_{p,\xi}(\{|X| \le t\}) = F_{p,\xi}^w(t) - F_{p,\xi}^w(-t) = H_{p,\xi}(t)$. We therefore proved the following.

**Lemma 6.1.** *The most dispersed symmetric probability measure in $\mathcal{F}_p^w(\xi)$ is $\nu_{p,\xi}$. This distribution achieves the equality $\nu_{p,\xi}(\{|X| \le t\}) = H_{p,\xi}(t) \le \nu(\{|X| \le t\})$ for all $\nu \in \mathcal{F}_p^w(\xi)$, and all $t > 0$.*

It turns out that this most dispersed distribution is also the least favorable distribution for soft thresholding. In order to see this fact, define the function $\mathsf{mse}_0 : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ by letting

$$\mathsf{mse}_0(x; \tau) = \mathbb{E}\{[\eta(x + Z; \tau) - x]^2\}, \tag{6.7}$$

whereby expectation is taken with respect to $Z \sim \mathsf{N}(0, 1)$. We then have the following useful calculus lemma (see for instance [DMM09].

**Lemma 6.2.** *For each $\tau \in [0, \infty)$, the mapping $x \mapsto \mathsf{mse}_0(x; \tau)$ is strictly monotone increasing in $x \in [0, \infty)$.*

Now the mean square error of scalar soft thresholding, cf. Eq. (2.5), is given by

$$\mathsf{mse}(1; \nu, \tau) \equiv \mathbb{E}\mathsf{mse}_0(|X|; \tau), \tag{6.8}$$

where expectation is taken with respect to $X \sim \nu$. From the above remarks, we obtain immediately the following characterization of the minimax problem.

**Corollary 6.1** (Saddlepoint). *Consider the game against Nature where the statistician chooses the threshold $\tau$, Nature chooses the distribution $\nu \in \mathcal{F}_p^w(\xi)$, and the statistician pays Nature an amount equal to the mean square error $\mathsf{mse}(1; \tau, \nu)$.*

*This game has a saddlepoint $(\tau_p^w(\xi), \nu_{p,\xi}^w)$, i.e. a pair satisfying*

$$\mathsf{mse}(1; \tau, \nu_{p,\xi}^w) \ge \mathsf{mse}(1; \tau_p^w(\xi), \nu_{p,\xi}^w) \ge \mathsf{mse}(1; \tau_p^w(\xi), \nu) \qquad \forall \tau > 0, . \tag{6.9}$$

*for all $\tau \ge 0$, and $\nu \in \mathcal{F}_p^w(\xi)$. In particular, the least-favorable probability measure is $\nu_{p,\xi}^w = \nu_{p,\xi}$, with distribution $F_{p,\xi}$ given in closed form by Eq. (6.6), and we have the following formula for the soft thresholding minimax risk:*

$$M_p^w(\xi) = \inf_{\tau \ge 0} \mathsf{mse}(1; \tau, \nu_{p,\xi}). \tag{6.10}$$
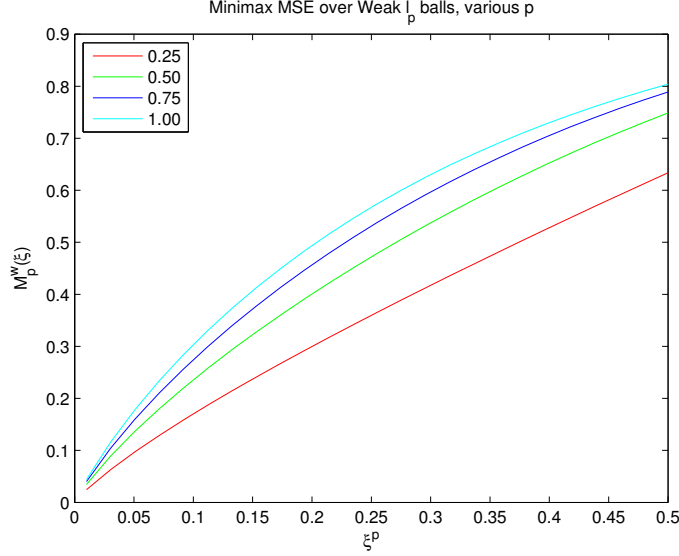
Figure 9: Minimax soft thresholding MSE over weak-$\ell_p$ balls, $M_p^w(\xi)$, for various $p$. Vertical axis: worst case MSE over $\mathcal{F}_p^w(\xi)$. Horizontal axis: $\xi^p$. Red, green, blue, aqua curves (from bottom to top) correspond to $p = 0.25, 0.50, 0.75, 1.00$.

Ordinarily, identifying a saddlepoint requires search over two variables, namely the threshold $\tau$ and the distribution $\nu$. In the present problem we need only search over one scalar variable, i.e. $\tau$. We can further make explicit the MSE calculation, by noting that, by Eq. (6.6)

$$\mathsf{mse}(1; \tau, \nu_{p,\xi}^w) = p \cdot \xi^p \int_\xi^\infty \mathsf{mse}_0(x; \tau) x^{-p-1}\, \mathrm{d}x\,. \tag{6.11}$$

By a simple calculus exercise, this formula and Lemma 6.2, imply the following.

**Lemma 6.3.** *The function $M_p^w(\xi)$ is strictly monotone increasing in $\xi \in (0, \infty)$. Hence, the inverse function*

$$(M_p^w)^{-1}(m) = \inf\left\{\, \xi : \; M_p^w(\xi) \geq m \right\},$$

*is well-defined for $m \in (0, 1)$.*

The asymptotic behavior of $M_p^w(\xi)$ in the very sparse limit $\xi \to 0$ was derived in [Joh93].

**Lemma 6.4** ([Joh93]). *As $\xi \to 0$, the minimax threshold level achieving Eq. (6.10) is given by*

$$\tau_p^w(\xi) \;\; = \;\; \sqrt{2\log(1/\xi^p)} \cdot \{1 + o(1)\}\,, ,$$

*and the corresponding minimax mean square error behaves, in the same limit, as*

$$M_p^w(\xi) = \frac{2}{2-p} \left(2\log(1/\xi^p)\right)^{1-p/2} \cdot \xi^p \cdot \{1 + o(1)\}\,. \tag{6.12}$$
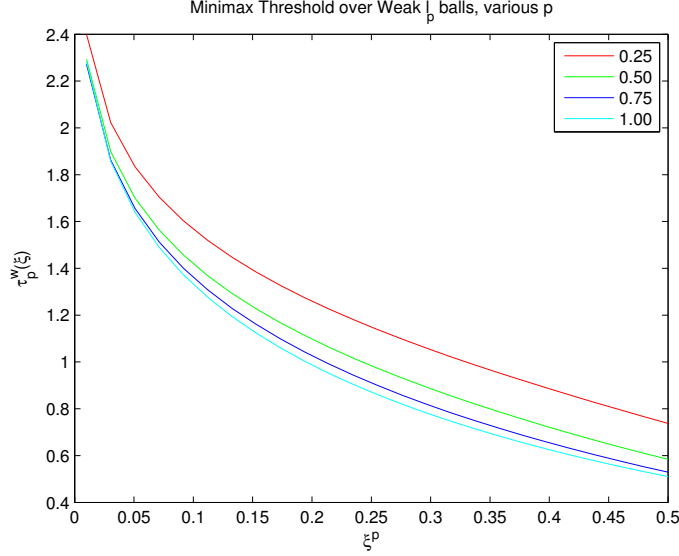
Figure 10: Minimax soft threshold parameter, $\tau_p^w(\xi)$, various $p$. Vertical axis: minimax threshold over $\mathcal{F}_p^w(\xi)$. Horizontal Axis: $\xi^p$. Red, green, blue, aqua curves (from top to bottom) correspond to $p = 0.25, 0.50, 0.75, 1.00$.

Comparing with Lemma 2.3, we see that the minimax threshold $\tau_p^w(\xi)$ coincides asymptotically with the one for strong $\ell_p$ balls. The corresponding risk is larger by a factor $2/(2-p)$ reflecting the larger set of possible distributions $\nu \in \mathcal{F}_p^w(\xi)$. For later use also note:

$$(M_p^w)^{-1}(m) = \left(\frac{2-p}{p}m\right)^{1/p} \cdot \left(2\log(\frac{2-p}{p}m)^{-1})\right)^{1/2-1/p} \cdot (1+o(1)), \qquad m \to 0.$$

## 6.2 Minimax MSE in Compressed Sensing under Weak $p$-th Moments

We return now to the compressed sensing setup. In the noiseless case we consider sequences of instances $\mathsf{S} \equiv \{I_{n,N}\} = \{(x_0^{(N)}, z^{(n)} = 0, A^{(n,N)})\}_{n,N}$ in $\mathcal{S}_p^w(\delta, \xi, 0)$. The minimax asymptotic mean square error of the LASSO is then given by considering the worst case sequence of instances

$$M_p^{w,*}(\delta, \xi) \equiv \sup_{\mathsf{S} \in \mathcal{S}_p^w(\delta, \xi, 0)} \inf_{\lambda \in \mathbb{R}_+} \mathrm{AMSE}(\lambda; \mathsf{S}). \tag{6.13}$$

Here asymptotic mean-square error is defined as per Eq. (1.4).

Analogously, in the noisy case $\sigma > 0$, we consider sequences of instances $\mathsf{S} \in \mathcal{S}_p^w(\delta, \xi, \sigma)$, We then define the minimax risk as

$$M_p^{w,*}(\delta, \xi, \sigma) \equiv \sup_{\mathsf{S} \in \mathcal{S}_p^w(\delta, \xi, \sigma)} \inf_{\lambda \in \mathbb{R}_+} \mathrm{AMSE}(\lambda; \mathsf{S}). \tag{6.14}$$

29

It turns out that complete analogs of the results of Sections 4 and 5 hold for the weak $p$-th moment setting. Since the proofs are easy modifications of the ones for strong $\ell_p$ balls, we omit them.

**Theorem 6.1** (Noiseless Case, Weak $p$-th moment). *For $\delta \in (0,1)$, $\xi > 0$, the Minimax AMSE of the LASSO over the weak-$\ell_p$ ball of radius $\xi$ is:*

$$M_p^{w,*}(\delta, \xi) = \frac{\delta \xi^2}{(M_p^w)^{-1}(\delta)^2} \tag{6.15}$$

*where $(M_p^w)^{-1}(\delta)$ is the inverse function of the soft thresholding minimax risk, see Eq. (6.10).*
    *Further we have:*
**Least Favorable $\nu$.** *The least-favorable distribution $\nu^{w,*}$ is the most dispersed distribution $\nu_{p,\xi}$ whose distribution function is given by Eq. (6.6), with $\xi = \xi^*$.*
**Minimax Threshold.** *The minimax threshold $\lambda^{w,*}(\delta, \xi)$ is given by the calibration relation (3.4) with $\tau = \tau^{w,*}(\delta, \xi)$ determined by:*

$$\tau^{w,*}(\delta, \xi) = \tau_p^w((M_p^w)^{-1}(\delta)), \tag{6.16}$$

*where $\tau_p^w(\cdot)$ is the soft thresholding minimax threshold, achieving the infimum in Eq. (6.10).*
**Saddlepoint.** *The pair $(\lambda^{w,*}, \nu^{w,*})$ satisfies a saddlepoint relation. Put for short $\mathrm{AMSE}(\lambda; \nu) = \mathrm{AMSE}(\lambda; \delta, \nu, \sigma = 0)$. The minimax AMSE is given by*

$$M_p^{w,*}(\delta; \xi) = \mathrm{AMSE}(\lambda^{w,*}; \nu^{w,*}),$$

*and*

$$
\begin{aligned}
\mathrm{AMSE}(\lambda^{w,*}; \nu^{w,*}) &\leq \mathrm{AMSE}(\lambda; \nu^{w,*}), &\forall \lambda > 0 \tag{6.17}\\
&\geq \mathrm{AMSE}(\lambda^{w,*}; \nu), &\forall \nu \in \mathcal{F}_p^w(\xi). \tag{6.18}
\end{aligned}
$$

As an illustration of this theorem, consider again the limit $\delta = n/N \to 0$ after $N \to \infty$ (equivalently, $n/N \to 0$ sufficiently slowly). It follows from Eq. (6.12) that

$$M_p^{w,*}(\delta, 1) = \left(1 - \frac{p}{2}\right)^{-2/p} \delta^{1-2/p}(2\log(\delta^{-1}))^{2/p-1}\{1 + o_\delta(1)\}, \qquad \delta \to 0.$$

We can also compute the minimax regularization parameter. Lemma 6.4 gives

$$\lambda^{w,*}(\delta, \xi) = \xi \cdot \left(1 - \frac{p}{2}\right)^{-1/p} \cdot \left(\frac{2\log(1/\delta)}{\delta}\right)^{1/p}\{1 + o(1)\}, \quad \delta \to 0. \tag{6.19}$$

In the noisy case, we get a result in many respects similar to the $p$th moment result.

**Theorem 6.2** (Noisy Case, Weak $p$-th moment). *For any $\delta, \xi > 0$, let $m^* = m_p^{w,*}(\delta, \xi)$ be the unique positive solution of*

$$\frac{m^*}{1 + m^*/\delta} = M_p^w\left(\frac{\xi}{(1 + m^*/\delta)^{1/2}}\right). \tag{6.20}$$

*Then the LASSO minimax mean square error $M_p^{w,*}$ is given by:*

$$M_p^{w,*}(\delta, \xi, \sigma) = \sigma^2 m_p^{w,*}(\delta, \xi/\sigma). \tag{6.21}$$

*Further, denoting by $\xi^* \equiv (1 + m^*/\delta)^{-1/2}\xi/\sigma$, we have:*

**Least Favorable $\nu$.** *The least-favorable distribution $\nu^{w,*}$ is the most dispersed distribution $\nu_{p,\xi}$ whose distribution function is given by Eq. (6.6).*

**Minimax Threshold.** *The minimax threshold $\lambda^*(\delta, \xi, \sigma)$ is given by the calibration relation (3.4) with $\tau = \tau^*(\delta, \xi, \sigma)$ determined as follows:*

$$\tau^{w,*}(\delta, \xi, \sigma) = \tau_p^w(\xi^*). \tag{6.22}$$

*where $\tau_p^w(\cdot)$ is the soft thresholding minimax threshold, achieving the infimum in Eq. (6.10).*

**Saddlepoint.** *The above quantities obey a saddlepoint relation. Put for short $\mathrm{AMSE}(\lambda; \nu) = \mathrm{AMSE}(\lambda; \delta, \nu, \sigma)$. The minimax AMSE obeys*

$$M_p^{w,*}(\delta, \xi) = \mathrm{AMSE}(\lambda^{w,*}; \nu^{w,*}),$$

*and*

$$
\begin{aligned}
\mathrm{AMSE}(\lambda^*; \nu^*) &\leq \mathrm{AMSE}(\lambda; \nu^{w,*}), && \forall \lambda > 0 & (6.23) \\
&\geq \mathrm{AMSE}(\lambda^{w,*}; \nu), && \forall \nu \in \mathcal{S}_p^w(\xi). & (6.24)
\end{aligned}
$$

# 7 Traditionally-scaled $\ell_p$-norm Constraints

This paper uses a non-traditional scaling $\|x_0\|_p^p \leq N \cdot \xi^p$ for the radius of $\ell_p$ balls; traditional scaling would be $\|x_0\|_p^p \leq \xi^p$. In this section we discuss the translation between the two types of conditions. We first define sequence classes based on norm constraints.

**Definition 7.1.** *The **traditionally-scaled $\ell_p$ problem suite** $\tilde{\mathcal{S}}_p(\delta, \xi, 0)$ is the class of sequences of problem instances $I_{n,N} = (x_0^{(N)}, z^{(n)}, A^{(n,N)})$ where:*
*(1) $n/N \to \delta$;*
*(2) $\|x_0^{(N)}\|_p^p \leq \xi^p$, and, for some sequence $B = \{B_M\}_{M \geq 0}$ such that $B_M \to 0$, we have $\sum_{i=1}^N (x_{0,i}^{(N)})^2 \mathbb{I}(|x_{0,i}^{(N)}| \geq M) \leq B_M N^{1-2/p}$ for every $N$;*
*(3) $z^{(n)} \in \mathbf{R}^n$, $\|z^{(n)}\|_2 \sim \sigma \cdot n^{1/2} \cdot N^{-1/p}$, $(n, N) \to \infty$.*
*(4) $A^{(n,N)} \sim \mathrm{GAUSS}(n, N)$.*

*The **traditionally-scaled weak $\ell_p$ problem suite** $\tilde{\mathcal{S}}_p^w(\delta, \xi, 0)$ is defined using conditions (1),(3),(4) and*
*($2^w$) $\|x_0^{(N)}\|_{w\ell_p}^p \leq \xi^p$, and, , for some sequence $B = \{B_M\}_{M \geq 0}$ such that $B_M \to 0$, we have $\sum_{i=1}^N (x_{0,i}^{(N)})^2 \mathbb{I}(|x_{0,i}^{(N)}| \geq M) \leq B_M N^{1-2/p}$ for every $N$;*

Comparing our earlier definitions of standard $\ell_p$-constrained problem suites $\mathcal{S}_p(\delta, \xi, \sigma)$ and $\mathcal{S}_p^w(\delta, \xi, \sigma)$ with these new definitions, conditions (1) and (4) are identical; while the new (2) and (3) are simply rescaled versions of corresponding conditions (2) and (3) in the earlier standard problem suites.[4] To deal with such rescaling, we need the following scale covariance property:

---

[4]Note the awkwardness of the noise scaling in the traditional scaling, as compared to the standard scaling used here

**Lemma 7.1.** *Let* $I = (x_0^{(N)}, z^{(n)}, A^{(n,N)})$ *be a problem instance and* $I_a = (a \cdot x_0^{(N)}, a \cdot z^{(n)}, A^{(n,N)})$ *be the corresponding dilated problem instance. Suppose that* $\hat{x}_\lambda^{(N)}$ *is the unique LASSO solution generated by instance* $I$ *and* $\hat{x}_\lambda^{(N),a}$ *the unique solution generated by instance* $I_a$. *Then*

$$\hat{x}_{a\lambda}^{(N),a} = a \cdot \hat{x}_\lambda^{(N)},$$

$$\|\hat{x}_{a\lambda}^{(N),a} - ax_0^{(N)}\|_2^2 = a^2 \cdot \|\hat{x}_\lambda^{(N)} - x_0^{(N)}\|_2^2,$$

*and*

$$\inf_\lambda E\|\hat{x}_\lambda^{(N),a} - ax_0^{(N)}\|_2^2 = a^2 \cdot \inf_\lambda E\|\hat{x}_\lambda^{(N)} - x_0^{(N)}\|_2^2.$$

Applying this lemma yields the following problem equivalences:

**Corollary 7.1.** *We have the scaling relations:*

$$\sup_{\mathsf{S} \in \tilde{\mathcal{S}}_p(\delta,\xi,0)} \inf_\lambda \mathrm{AMSE}(\lambda, \mathsf{S}) = N^{-2/p} \cdot \sup_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,0)} \inf_\lambda \mathrm{AMSE}(\lambda, \mathsf{S});$$

*and*

$$\sup_{\mathsf{S} \in \tilde{\mathcal{S}}_p^w(\delta,\xi,0)} \inf_\lambda \mathrm{AMSE}(\lambda, \mathsf{S}) = N^{-2/p} \cdot \sup_{\mathsf{S} \in \mathcal{S}_p^w(\delta,\xi,0)} \inf_\lambda \mathrm{AMSE}(\lambda, \mathsf{S}).$$

Let's apply this to noiseless $\ell_p$ ball constraint. By Theorem 4.1 we have

$$\min_\lambda \max_{\mathsf{S} \in \mathcal{S}(\delta,\xi,0)} \mathrm{AMSE}(\lambda, \mathsf{S}) = \frac{\delta \xi^2}{M_p^{-1}(\delta)^2}$$

Considering the *un*normalized squared error $\|\hat{x}_\lambda - x_0\|^2$ and operating purely formally, define a symbol $\bar{E}$ so that when $x_0^{(N)}$ arises from a given sequence $\mathsf{S}$,

$$\bar{E}\|\hat{x}_\lambda^{(N)} - x_0^{(N)}\|^2 = N \cdot \mathrm{AMSE}(\lambda, \mathsf{S}).$$

Remembering $\delta = n/N$ we have

$$\min_\lambda \max_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,0)} \bar{E}\|\hat{x}_\lambda^{(N)} - x_0^{(N)}\|^2 = N \cdot (n/N)^{1-2/p} \cdot \xi^2 \cdot (2\log(N/n))^{2/p-1} \{1 + o_N(1)\}.$$

$$= N^{2/p}\xi^2 \cdot \left(\frac{2\log(N/n)}{n}\right)^{2/p-1} \{1 + o_N(1)\}.$$

Using the traditionally-scaled $\ell_p$ problem suite,

$$\min_\lambda \max_{\mathsf{S} \in \tilde{\mathcal{S}}_p(\delta,\xi,0)} \bar{E}\|\hat{x}_\lambda - x_0\|^2 = N^{-2/p} \cdot \min_\lambda \max_{\mathsf{S} \in \mathcal{S}_p(\delta,\xi,0)} \bar{E}\|\hat{x}_\lambda - x_0\|^2,$$

where on the LHS we have $\tilde{\mathcal{S}}_p(\delta, \xi, 0)$ while on the RHS we have $\mathcal{S}_p(\delta, \xi, 0)$. We conclude

**Corollary 7.2.** *Consider the noiseless, traditionally-scaled* $\ell_p$ *problem formulation. The asymptotic MSE for the* $\ell^2$*-norm error measure has the asymptotic form*

$$\min_\lambda \max_{\mathsf{S} \in \tilde{\mathcal{S}}_p(\delta,\xi,0)} \bar{E}\|\hat{x}_\lambda - x_0\|^2 = \xi^2 \cdot \left(\frac{2\log(N/n)}{n}\right)^{2/p-1} \{1 + o_N(1)\}; \qquad (7.1)$$

*this is valid both for $n/N \to \delta \in (0,1)$ and for $\delta = n/N \to 0$ slowly enough. The maximin penalization has an elegant prescription when $n/N \to 0$ slowly enough:*

$$\lambda^* = \xi \cdot \left( \frac{2 \log(N/n)}{n} \right)^{1/p}, \qquad \mathsf{S} \in \tilde{\mathcal{S}}_p(\delta, \xi, 0). \tag{7.2}$$

Our results can now be compared with earlier results written in the traditional scaling. We rewrite our result for $\xi = 1$, using a simple moment condition that implies uniform integrability. For all sufficiently large $B$, and all $q > 2$, we obtained:

$$\min_{\lambda} \max_{\|x_0\|_p^p \leq 1,\, \|x_0\|_q^q \leq BN^{1-q/p}} \bar{E} \|\widehat{x}_\lambda^{(N)} - x_0^{(N)}\|^2 = \left( \frac{2 \log(N/n)}{n} \right)^{2/p-1} \left\{ 1 + o_N(1) \right\}. \tag{7.3}$$

In the case $\lambda = 0$, earlier results [Don06a, CT05] imply:

$$\max_{\|x_0\|_p^p \leq 1} \|\widehat{x}_0 - x_0\|^2 = O_P\left( \left( \frac{\log(N/n)}{n} \right)^{2/p-1} \right), \qquad n/N \to 0. \tag{7.4}$$

There are two main differences in *technical content* between the new result and earlier ones

- The use of $\bar{E}$ on the LHS of (7.3) versus $O_P(\,\cdot\,)$ on the RHS of (7.4).

- The supremum over $\{\|x_0^{(N)}\|_p^p \leq 1\}$ on the LHS of (7.4) versus the supremum over $\{\|x_0^{(N)}\|_p^p \leq 1\} \cap \{\|x_0\|_q^q \leq BN^{1-q/p}\}$ on the LHS of (7.3).

The main difference in *results* is of course that the new result gives a precise constant in place of the $O(\,\cdot\,)$ result which was previously known. See Section 10.3 for further discussion.

The new result has the additional ingredient, not seen earlier, that we constrain not only $\{\|x_0^{(N)}\|_p^p \leq 1\}$ but also $\{\|x_0^{(N)}\|_2^2 \leq BN^{1-q/p}\}$. For each $p < 2$, this additional constraint does indeed give a smaller set of feasible vectors for large $N$. See Section 10.2 for further discussion.

A traditionally-scaled weak-$\ell_p$ problem suite $\tilde{\mathcal{S}}_p^w(\delta, \xi, \sigma)$ can also be defined; without giving details, we have:

**Corollary 7.3.** *Consider the noiseless, traditionally-scaled weak-$\ell_p$ problem formulation. The asymptotic MSE for the $\ell^2$-norm error measure has the asymptotic form*

$$\min_{\lambda} \max_{\mathsf{S} \in \tilde{\mathcal{S}}_p^w(\delta, \xi, 0)} \bar{E} \|\widehat{x}_\lambda - x_0\|^2 = (1 - p/2)^{-2/p} \cdot \xi^2 \cdot \left( \frac{2 \log(N/n)}{n} \right)^{2/p-1} \left\{ 1 + o_N(1) \right\}; \tag{7.5}$$

*this is valid both for $n/N \to \delta \in (0,1)$ and for $\delta = n/N \to 0$ slowly enough. The maximin penalization has an elegant prescription for $n/N$ small:*

$$\lambda^* = (1 - p/2)^{-1/p} \cdot \xi \cdot \left( \frac{2 \log(N/n)}{n} \right)^{1/p}, \qquad \mathsf{S} \in \tilde{\mathcal{S}}_p^w(\delta, \xi, 0). \tag{7.6}$$

# 8   Compressed Sensing over the Bump Algebra

Our discussion involving $\ell_p$-balls is so far rather abstract. We consider here a stylized application: recovering a signal $f$ in the Bump Algebra from compressed measurements. Consider a function $f : [0, 1] \to \mathbb{R}$ which admits the representation

$$f(t) = \sum_{i=1}^{\infty} c_i\, g\big((t - t_i)/\sigma_i\big), \qquad g(x) = \exp\big(-x^2/2\big), \quad \sigma_i > 0. \tag{8.1}$$

Each term $g(\,\cdot\,)$ is a Gaussian 'bump' normalized to height 1, and we assume $\sum_{i=1}^{\infty} |c_i| \leq 1$ which ensures convergence of the series. The $c_i$ are signed amplitudes of the 'bumps' in $f$. We refer to the book by Yves Meyer [Mey84] and also to the discussion in [DJ98], which calls such objects models of *polarized spectra*. Any such function also has a wavelet representation

$$f = \sum_{j \geq -1} \sum_{k \in \mathcal{I}_j} \alpha_{j,k} \psi_{j,k},$$

where the $\psi_{j,k}$ are smooth orthonormal wavelets (for example Meyer wavelets or Daubechies wavelets), and the wavelet coefficients obey $\sum_{j,k} |\alpha_{j,k}| \leq C$. The constant $C$ depends only on the wavelet basis [Mey84]. Here $j$ denotes the level index, and $k$ the position index. We have $|\mathcal{I}_{-1}| = 1$, and $|\mathcal{I}_j| = 2^j$ for each $j \geq 0$. In other words the collection of functions with wavelet coefficients in an $\ell_1$-ball of radius $C$ contains the whole algebra of functions representable as in (8.1).

Now consider compressed sensing of such an object. We fix a maximum resolution, by picking $N = 2^J$ and considering the finite-dimensional problem of recovering the object $f_N = \sum_{j<J} \sum_{k=0}^{2^j-1} \alpha_{j,k} \psi_{j,k}$. The scale $2^{-J}$ corresponds to an effective discretization scale: on intervals of length much smaller than $2^{-J}$, the function $f_N$ is approximately constant. Reconstructing the function $f_N$ is equivalent to recovering the $2^J$ coefficients

$$x_0 = \big(\alpha_{-1,0}, \alpha_{0,0}, \alpha_{1,0}, \alpha_{1,1}, \alpha_{2,0}, \ldots, \alpha_{2,3}, \alpha_{3,1}, \ldots, \alpha_{J-1,0}, \ldots, \alpha_{J-1,2^{J-1}-1}\big).$$

We know that coefficients at scales 1 through $J - 1$ combined have a total $\ell_1$-norm bounded by a numerical constant $C$. Without loss of generality, we shall take $C = 1$ (this corresponds to rescaling the constraint on the bump representation (8.1)).

Denote by $V_J$ the $2^J$-dimensional space of functions on $[0, 1]$, with resolution $2^{-J}$, i.e.

$$V_J \equiv \left\{ \sum_{j<J} \sum_{k=0}^{2^j-1} \alpha_{j,k} \psi_{j,k} : \ \alpha_{j,k} \in \mathbb{R} \right\}. \tag{8.2}$$

We can construct a random linear *measurement operator* $\mathcal{A} : V_J \to \mathbb{R}^n$, such that the matrix $A$ representing $\mathcal{A}$ in the basis of wavelets has random Gaussian coefficients iid $\mathsf{N}(0, 1)$. We then take $n+1$ noiseless measurements: the scalar $\alpha_{-1,0} = \langle f, \psi_{-1,0}\rangle$ associated to the 'father wavelet', and the vector $y = \mathcal{A} f_N$. Notice that, since the measurements are noiseless, the variance of the entries of the measurement matrix $A$ can be rescaled arbitrarily.

In the wavelet basis, the measurements can be rewritten as $y = Ax_0$, where the $A$ is an $n \times N$ Gaussian random matrix. This is precisely a problem of the type studied in earlier sections. Suppose now that we apply $\ell_1$-penalized least-squares

$$\widehat{x}_\lambda \equiv \arg \min_x \left\{ \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 \right\}, \tag{8.3}$$

and denote the entries of the reconstruction vector by $\widehat{x}_\lambda \equiv (\widehat{\alpha}_{0,0}, \ldots, \widehat{\alpha}_{J-1,2^{J-1}-1})$. The function $f_N$ is therefore reconstructed as $\widehat{f}_N$, where

$$\widehat{f}_N = \sum_{j<J} \sum_{k=0}^{2^j-1} \widehat{\alpha}_{j,k} \psi_{j,k} \,.$$

We adopt the performance measure

$$\mathrm{MSE}(\widehat{f}_N, f_N) \equiv \mathbb{E}\big\{\|f_N - \widehat{f}_N\|_{L_2[0,1]}^2\big\} = \mathbb{E}\|x_0 - \widehat{x}_\lambda\|_2^2;$$

where the last equality uses the orthonormality of the wavelet basis.

We wish to choose an appropriate value of $\lambda \geq 0$ to give the best reconstruction performance. Note that the coefficients vector $x_0 \in \mathbb{R}^N$ satisfies by assumption

$$\|x_0\|_1 \leq 1$$

so we are in the setting of traditionally-scaled $\ell_p$ balls. The discussion of the last section now applies; we obtain results by rescaling results from Theorem 4.1. Letting $\lambda_p^*(\delta, \xi)$ denote the minimax threshold of Theorem 4.1, define

$$\lambda_N = N^{-1} \cdot \lambda_1^*(\frac{n}{N}, 1). \tag{8.4}$$

**Corollary 8.1.** *Consider a sequence of functions $f_N \in V_J$ in the Bump Algebra (normed so that the wavelet coefficients have $\ell_1$-norm bounded by 1). Consider Gaussian measurement operators $\mathcal{A}_N : V_J \to \mathbb{R}^n$ indexed by the problem dimensions $N = 2^J$, and $n$. Let $\widehat{f}_N^*$ denote the reconstruction of $f_N$ using regularization parameter $\lambda = \lambda_N$ of (8.4).*

*(i) Assume $n/N \to \delta \in (0, 1)$. Then we have*

$$\mathrm{MSE}(\widehat{f}_N^*, f_N) \leq N^{-1} \cdot M_1^*(\delta, 1)\,(1 + o(1))\,, \tag{8.5}$$

*with $M_1^*(\delta, \xi)$ as in Theorem 4.1. This bound is asymptotically tight (achieved for a specific sequence $f_N$).*

*(ii) Assume $n/N \to 0$ sufficiently slowly. Then we have*

$$\mathrm{MSE}(\widehat{f}_N^*, f_N) \leq \frac{2\log(N/n)}{n} \cdot (1 + o(1))\,, \, \tag{8.6}$$

*and the bound is asymptotically tight (achieved for a specific sequence $f_N$).*

# 9 Compressed Sensing over Bounded Variation Classes

Compressed sensing problems make sense for many other functional classes. The class of Bounded Variation affords an application of our results on weak $\ell_p$ classes.

1. Every bounded variation function $f \in BV[0, 1]$ has Haar wavelet coefficients in a weak-$\ell_{2/5}$ ball.

2. Every $f \in BV[0, 1]^2$ has wavelet coefficients in a weak-$\ell_1$ ball [CDPX99].

We can develop a theory of compressed sensing over BV spaces following the previous section, now using Haar wavelets. $V_J$ means again the span wavelets of spatial scale $2^{-J}$ or coarser. We let $d$ denote the spatial dimension ($d = 1$ or $2$ in the above examples). We use regularization parameter

$$\lambda_N = N^{-1} \cdot \lambda^{w,*}(n/N, 1). \tag{9.1}$$

**Corollary 9.1.** *Consider a sequence of functions $f_N \in V_J$ whose Haar wavelet coefficients have weak $\ell_p$-norm bounded by $1$. Consider Gaussian measurement operators $\mathcal{A}_N : V_J \to \mathbb{R}^n$ indexed by the problem dimensions $N = 2^{dJ}$, and $n$. Let $\widehat{f}_N^*$ denote the reconstruction of $f_N$ using regularization parameter $\lambda = \lambda_N$ of (9.1).*

*(i) Assume $n/N \to \delta \in (0, 1)$. Then we have*

$$\mathrm{MSE}(\widehat{f}_N^*, f_N) \le N^{-1} \cdot M_p^{w,*}(\delta, 1)\,(1 + o(1))\,, \tag{9.2}$$

*with $M_1^{*,w}(\delta, \xi)$ as in Theorem 6.1. This bound is asymptotically tight (achieved for a specific sequence $f_N$).*

*(ii) Assume $n/N \to 0$ sufficiently slowly. Then we have*

$$\mathrm{MSE}(\widehat{f}_N^*, f_N) \le \left(1 - \frac{p}{2}\right)^{-2/p} \cdot \left(\frac{2\log(N/n)}{n}\right)^{2/p-1} \cdot (1 + o(1)), \, , \tag{9.3}$$

*and the bound is asymptotically tight (achieved for a specific sequence $f_N$).*

Although BV offers only the applications $p = 1$ ($d = 2$) and $p = 2/5$ ($d = 1$), weak-$\ell_p$ spaces arise elsewhere, and serve as useful models for image content. For example, for images containing smooth edges, we have the following model: every $f : [0, 1]^2 \mapsto \mathbb{R}$ which is locally in $C^2$ except at $C^2$ 'edges' has curvelet coefficients levelwise in weak-$\ell_{2/3}$ balls [CD04]. Our compressed sensing result for BV can be adapted without change to the conclusions for such a setting, after replacing the role of Haar wavelets by Curvelets.

## 10 Discussion

In this last section we discuss some specific aspects of our results and overview (in an unavoidably incomplete way) the related literature.

### 10.1 Equivalence of Random and Deterministic Signals/Noises

A striking aspect of our results is the equivalence of random and deterministic signals and noises (traceable here to Proposition 3.1). The AMSE formula in the general case, as given by Eq. (3.5), depends on the sequence of signals $x_0^{(N)}$ and of noise vectors $z^{(n)}$ only through simple statistics of such vectors. More precisely, it depends only on their asymptotic empirical distributions, respectively $\nu$ and $\omega$. In fact the dependence on $z^{(n)}$ is even weaker: the asymptotic risk only depends on the limit second moment $\mathbb{E}_\omega(Z^2)$.

At first sight, these findings are somewhat surprising. For instance we might replace $x_0^{(N)}$ with a random vector with i.i.d. entries with common distribution $\nu$ without changing the asymptotic risk. This asymptotic equivalence between random and deterministic signal is in fact a quite simple and robust consequence of the absence of structure of the measurement matrix $A$. We do not spell out the details here, but note the following simple facts

1. Under our model for $A$, the columns of $A$ are exchangeable, so there is no distributional difference between $Ax_0$ and $APx_0$, for any permutation matrix $P$.

2. As a consequence, there is no difference in expected performance between a fixed vector $x_0$ and a random vector obtained by permuting the entries of $x_0$ uniformly at random.

3. Asymptotically for large $N$ there is a negligible difference in performance between a fixed vector $x_0$ and the typical random vector obtained by sampling with replacement from the entries of $x_0$.

This argument implies that we can replace the deterministic vectors $x_0^{(N)}$ with random vectors with i.i.d. entries. As the argument clarifies, this phenomenon ought to exist for more general models of $A$.

## 10.2 Comparison with Previous Approaches

Much of the analysis of compressed sensing reconstruction methods has relied so far on a kind of *qualitative analysis*. A typical approach has been to frame the analysis in terms of 'worst case' conditions on the measurement matrix $A$. A useful set of conditionsis provided by the restricted isometry property (RIP), [CT05, CRT06] and refinements [BRT09, vdGB09, BGI$^+$08]. These conditions are typically pessimistic, in that they assume that the signal $x_0$ is chosen adversarially, but they capture the correct scaling behavior.

The advantage of this approach is its broad applicability; since one assumes little about the matrix $A$, the derived bound will perhaps apply to a wide range of matrices. However, there are two limitations:

(a) These conditions have been proved to hold with linear scaling of $\|x_0\|$ and $n$ with the signal dimension $N$, only for specific random ensembles of measurement matrices, e.g. random matrices with i.i.d. subexponential entries.

(b) The resulting bounds typically only hold up to unspecified numerical constants. Efforts to make precise the implied constants in specific cases (see for instance [BCT11]) show that this approach imposes restrictive conditions on the signal sparsity. For instance, for a Gaussian measurement matrix with undersampling ratio $\delta = 0.1$, RIP implies successful reconstruction [BCT11] only if $\|x_0\|_0 \lesssim 0.0002\,N$. In empirical studies, a much larger support appears to be tolerated.

The present paper works with only one matrix ensemble – Gaussian random matrices – but gets quantitatively precise results, like the companion works [DMM09, DMM10, BM10]. The approach provides sharp performance guarantees under suitable probabilistic models for the measurement process.

To be concrete, consider the case of $x_0$ belonging to the weak-$\ell_p$ ball of radius 1, $\|x_0\|_{w\ell_p} \leq 1$. Building on the RIP theory, the review paper [Can06] derives the bound

$$\|\widehat{x}_\lambda - x_0\|^2 \;\; \leq \;\; C\left(\frac{\log(N/n)}{n}\right)^{2/p-1}, \tag{10.1}$$

holding for Gaussian measurement matrices $A$, and for unspecified constant $C$. Analogous minimax bounds for $\ell_p$ balls are known [Don06a, RWY09]. Our results have the same form, but with specific constants, e.g. $C = (1 - (p/2))^{-2/p}$ for weak-$\ell_p$ balls, cf. Eq. (7.5) and $C = 1$ for ordinary $\ell_p$ balls, cf. Eq. (7.1). Moreover, these constants are sharp, i.e. attained by specifically described $x_0$.

Let us finally mention the recent paper [CP10], that takes a probabilistic point of view similar to the one of [DMM10] and to the present one, although using different techniques. This approach avoids using RIP or similar conditions, and applies to a broad family of matrices with i.i.d. rows. On the other hand, it only allows to prove upper bounds on MSE off by logarithmic factors.

## 10.3  Comparison to the theory of widths

Recall that the Gel'fand $n$-width of of a set $K \subseteq \mathbb{R}^N$ with respect to the norm $\| \cdot \|_X$ is defined as

$$d_n(K, X) = \inf_{A \in \mathbb{R}^{n \times N}} \sup_{x \in K \cap \ker(A)} \|x\|_X \,, \tag{10.2}$$

where $\ker(A) \equiv \{v \in \mathbb{R}^N : Av = 0\}$. Here we shall consider $K$ to be the $\ell_p$ ball of radius 1, $B_p^N \equiv \{x \in \mathbb{R}^N : \|x\|_p \le 1\}$, and fix $\| \cdot \|_X$ to be the ordinary $\ell_2$ norm. A series of works [Kas77, GG84, Don06a, FPRU10] established that

$$d_n(B_p^N, \ell_2) \ge c_p \left( \frac{\log(N/n)}{n} \right)^{1/p - 1/2} \tag{10.3}$$

as long as the term in parenthesis is smaller than 1.

The interest for us lies in the well-known observation [Don06a] that $d_n(B_p^N, \ell_2)$ provides a lower bound on the compressed sensing mean square error under arbitrary reconstruction algorithm, and for arbitrary measurement matrix $A$. In particular

$$\max_{x_0 \in B_p^N} \|\widehat{x} - x_0\| \ge d_n(B_p^N, \ell_2) \,. \tag{10.4}$$

So it makes sense to define the inefficiency of a certain matrix/reconstruction procedure as the ratio of the two sides in the above inequality

$$r_{\mathrm{alg}}(B_p^N, \ell_2) \equiv \frac{1}{d_n(B_p^N, \ell_2)} \max_{x_0 \in B_p^N} \|\widehat{x} - x_0\| \,. \tag{10.5}$$

This ratio implicitly depends on the matrix $A$. In the case $p = 1$, $\lambda = 0$ it is known that $\ell^1$ minimization is inefficient at most by a factor 2:

$$r_{\min \ell_1}(B_p^N, \ell_2) \le 2; \tag{10.6}$$

(for example [Don06a] showed this by invoking [TW80]).

Our work concerns random Gaussian matrices and LASSO reconstruction. Since the worst-case performance of the optimally-tuned LASSO can not be worse than the worst-case performance of min-$\ell^1$ reconstruction, and since we have a formal expression for the

worst-case AMSE of optimally-tuned LASSO, the asymptotic formula (7.1) together with the bound (10.3) implies for all sufficiently large $B$ and any $q > 2$ that

$$\max_{x_0 \in B_p^N,\, \|x_0\|_q \leq BN^{1/q-1/p}} \overline{E} \|\widehat{x}_0^{(N)} - x_0\| = \sqrt{\frac{2\log(N/n)}{n}}(1 + o(1)), \tag{10.7}$$

with $\overline{E}$ defined in analogy with Section 7 The constant $B$ is arbitrary, which suggests (but of course does not prove) that we can remove the hypothesis $\|x_0\|_q \leq BN^{1/q-1/p}$ completely.

On the other hand, for a fixed matrix $A \in \mathbb{R}^{n \times N}$, we can define the width

$$d_n(K, A, X) = \sup_{x \in K \cap \ker(A)} \|x\|_X,$$

so that the Gel'fand $n$-width is the infimum of this quantity over $A$. Using results of Donoho and Tanner [JAMS, 2009] one can give the lower bound for $p = 1$ and Gaussian random matrices

$$d_n(B_p^N, A, \ell_2) \geq \sqrt{\frac{\log(N/n)}{4en}}(1 + o(1)).$$

The RHS is quantitatively quite close to the RHS of the right-side display, within a factor 2.5. Hence the results of this paper suggest that statistical methods may also provide geometric information.

In the general case $0 < p < 1$, lower bounds on $c_p$ are given in [FPRU10], but they do not appear as tight as desirable.

## 10.4 About the Uniform Integrability Condition

We have just seen once again that our hypotheses on $\ell_p$ balls can be scaled to match $\|x_0\|_p \leq 1$ but then they also include the hypothesis $\|x_0\|_q \leq BN^{1/q-1/p}$. It may seem at first glance that this is a serious additional constraint; it implies that the entries in $x_0$ cannot be very large as $N$ increases, whereas the condition $\|x_0\|_p \leq 1$ of course permits entries as large as 1.

However, note that our analysis identifies the least-favorable $x_0$, and that the constant $B$ plays no role. In fact, if we make a homotopy between the least-favorable object and objects requiring larger $B$, we find that the AMSE is decreasing in the direction of larger $B$. Pushing things to the extreme where $B$ goes unbounded, of course our analysis techniques no longer rigorously apply, but it is quite clear that this is an unpromising direction to move. Hence we believe that this is largely a technical condition, caused by our method of proof.

# References

[BCT11]  J. D. Blanchard, C. Cartis, and J. Tanner, *Compressed sensing: How sharp is the restricted isometry property?*, SIAM Review **53** (2011), 105–125.

[BGI+08]  R. Berinde, A.C. Gilbert, P. Indyk, H. Karloff, and M.J. Strauss, *Combining geometry and combinatorics: A unified approach to sparse signal recovery*, 47th Annual Allerton Conference (Monticello, IL), September 2008, pp. 798 – 805.

[BM10]    M. Bayati and A. Montanari, *The LASSO risk for gaussian matrices*, arXiv:1008.2581, 2010.

[BM11]    ———, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. on Inform. Theory **57** (2011), 764–785.

[BRT09]   P. J. Bickel, Y. Ritov, and A. B. Tsybakov, *Simultaneous analysis of Lasso and Dantzig selector*, Amer. J. of Mathematics **37** (2009), 1705–1732.

[Can06]   E. Candès, *Compressive sampling*, Proceedings of the International Congress of Mathematicians (Madrid, Spain), 2006.

[CD95]    S.S. Chen and D.L. Donoho, *Examples of basis pursuit*, Proceedings of Wavelet Applications in Signal and Image Processing III (San Diego, CA), 1995.

[CD04]    E.J. Candès and DL Donoho, *New tight frames of curvelets and optimal representations of objects with piecewise $C_2$ singularities*, Comm. Pure Appl. Math **57** (2004), 219–266.

[CDPX99]  A. Cohen, R. DeVore, P. Petrushev, and H. Xu, *Nonlinear Approximation and the Space* $BV(\mathbb{R}^2)$, Amer. J. of Mathematics **121** (1999), 587–628.

[CP10]    E. Candés and Y. Plan, *A probabilistic and RIPless theory of compressed sensing*, arXiv:1011.3854, 2010.

[CRT06]   E. Candes, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics **59** (2006), 1207–1223.

[CT05]    E. J. Candes and T. Tao, *Decoding by linear programming*, IEEE Trans. on Inform. Theory **51** (2005), 4203–4215.

[DJ94]    D. L. Donoho and I. M. Johnstone, *Minimax risk over $l_p$ balls*, Prob. Th. and Rel. Fields **99** (1994), 277–303.

[DJ98]    ———, *Minimax estimation via wavelet shrinkage*, Annals of Statistics **26** (1998), 879–921.

[DLM90]   D.L. Donoho, R.C. Liu, and K.B. MacGibbon, *Minimax risk over hyperrectangles, and implications*, Annals of Statistics **18** (1990), 1416–1437.

[DMM09]   D. L. Donoho, A. Maleki, and A. Montanari, *Message Passing Algorithms for Compressed Sensing*, Proceedings of the National Academy of Sciences **106** (2009), 18914–18919.

[DMM10]   D.L. Donoho, A. Maleki, and A. Montanari, *The Noise Sensitivity Phase Transition in Compressed Sensing*, arXiv:1004.1218, 2010.

[Don06a]  D. L. Donoho, *Compressed sensing*, IEEE Transactions on Information Theory **52** (2006), 489–509.

[Don06b]    D. L. Donoho, *High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension*, Discrete and Computational Geometry **35** (2006), no. 4, 617–652.

[DT05]    D. L. Donoho and J. Tanner, *Neighborliness of randomly-projected simplices in high dimensions*, Proceedings of the National Academy of Sciences **102** (2005), no. 27, 9452–9457.

[DT10]    ———, *Precise undersampling theorems*, Proc. IEEE (2010), to appear.

[DTDS06]    D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, *Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit*, Stanford Tecnical Report, 2006.

[FPRU10]    S. Foucart, A. Pajor, H. Rauhut, and T. Ullrich, *The Gelfand widths of $\ell_p$-balls for $0 < p \le 1$*, Journal of Complexity **26** (2010), 629–640.

[GG84]    A.Y. Garnaev and E.D. Gluskin, *On widths of the euclidean ball*, Soviet Mathematics Doklady **30** (1984), 200–203.

[Joh93]    I.M. Johnstone, *Minimax bayes, asymptotic minimax and sparse wavelet priors*, Statistical Decision Theory and Related Topics V, Springer-Verlag, 1993, pp. 303–326.

[Kas77]    B. Kashin, *Diameters of Some Finite-Dimensional Sets and Classes of Smooth Functions*, Math. USSR Izv. **11** (1977), 317–333.

[LDSP08]    M. Lustig, D.L Donoho, J.M. Santos, and J.M Pauly, *Compressed sensing mri*, IEEE Signal Processing Magazine (2008).

[Mey84]    Y. Meyer, *Wavelets*, Cambridge University Press, 1984.

[RWY09]    G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax rates of estimation for high-dimensional linear regression over $\ell_q$-balls*, 47th Annual Allerton Conference (Monticello, IL), September 2009.

[Tib96]    R. Tibshirani, *Regression shrinkage and selection with the lasso*, J. Royal. Statist. Soc B **58** (1996), 267–288.

[TW80]    J. F. Traub and H. Wozniakowski, *A general theory of optimal algorithms*, Academic Press, New York, 1980.

[vdGB09]    S. A. van de Geer and P. Bühlmann, *On the conditions used to prove oracle results for the Lasso*, Electron. J. Statist. **3** (2009), 1360–1392.