Universality of the LASSO Cost*

Phan-Minh Nguyen†

October 1, 2017

Abstract

Consider the linear inverse problem of reconstructing a vector $\boldsymbol{x}_0 \in \mathbb{R}^n$ from a noisy linear observation $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is random with independent and identically distributed entries, using the LASSO, which is the following optimization problem:

$$\mathrm{OPT}\left(\boldsymbol{A}\right) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{n}\left\{\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1\right\}.$$

Consider the asymptotic regime $n \to \infty$ and $m/n \to \delta > 0$, $\delta \neq 1$. For a given fixed $\lambda > 0$, we show that universality (with respect to the randomness of $\boldsymbol{A}$) holds for the LASSO cost $\mathrm{OPT}\left(\boldsymbol{A}\right)$. As an intermediate step in the proof, we obtain an extension of Kashin's theorem, which could be of independent interests.

1 Statement of the Result

Consider the linear model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}_0 + \boldsymbol{w}$, where $\boldsymbol{x}_0 \in \mathbb{R}^n$ is a vector to be reconstructed from the observation $\boldsymbol{y}$, $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the (known) sensing matrix, and $\boldsymbol{w} \in \mathbb{R}^m$ is the (unknown) noise. When $m < n$, the problem is underdetermined, and is generally encountered in compressed sensing and high-dimensional statistics. A common approach to this problem is to consider the LASSO:

$$\mathrm{OPT}\left(\boldsymbol{A}\right) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{n}\left\{\frac{1}{2}\left\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{y}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1\right\} \tag{1}$$

$$= \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{n} C\left(\boldsymbol{x}, \boldsymbol{A}\right), \tag{2}$$

in which

$$C\left(\boldsymbol{x}, \boldsymbol{A}\right) = \frac{1}{2}\left\|\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{x}_0\right) - \boldsymbol{w}\right\|_2^2 + \lambda\left\|\boldsymbol{x}\right\|_1 \tag{3}$$

for a given pre-chosen parameter $\lambda > 0$. To lighten the notation, we have dropped the dependence on $\boldsymbol{x}_0$, $\boldsymbol{w}$ and λ in $C\left(\boldsymbol{x}, \boldsymbol{A}\right)$ and $\mathrm{OPT}\left(\boldsymbol{A}\right)$. We focus on the asymptotic regime $n \to \infty$, thinking of the above as a sequence (in n) of problem instances. The main purpose of this note is to establish a universality property of $\mathrm{OPT}\left(\boldsymbol{A}\right)$, stated in the following.

Theorem 1. *Assume the following setting:*

*The result was proven during the preparation of the work [MN17], jointly by Andrea Montanari and the author of this note.

†Department of Electrical Engineering, Stanford University

- $\boldsymbol{x}_0 = \boldsymbol{x}_0(n)$ *is drawn from the sequence* $\{\boldsymbol{x}_0(n)\}_{n\in\mathbb{N}}$ *such that the empirical distribution of* $\boldsymbol{x}_0(n)$ *converges weakly to a probability measure associated with a random variable* $X_0 \in \mathbb{R}$, $\|\boldsymbol{x}_0(n)\|_2^2/n \to \mathbb{E}[X_0^2] \equiv \mathsf{M}_2 < \infty$, $\|\boldsymbol{x}_0(n)\|_1/n \to \mathbb{E}[|X_0|] \equiv \mathsf{M}_1 < \infty$, *and* $\|\boldsymbol{x}_0(n)\|_{10}^{10}/n \to \mathbb{E}[X_0^{10}] \equiv \mathsf{M}_{10} < \infty$.
- $\boldsymbol{A} = \boldsymbol{A}(n) \in \mathbb{R}^{m\times n}$ *for* $m = m(n)$ *that satisfies* $m/n \to \delta > 0$, $\delta \neq 1$, *and entries of* $\boldsymbol{A}(n)$ *are independent and identically distributed (i.i.d.), and*
- $\boldsymbol{w} = \boldsymbol{w}(n)$ *is drawn from the sequence* $\{\boldsymbol{w}(n)\}_{n\in\mathbb{N}}$ *such that the empirical distribution of* $\boldsymbol{w}(n)$ *converges weakly to a probability measure associated with a random variable* $W \in \mathbb{R}$, $\|\boldsymbol{w}(n)\|_2^2/m \to \mathbb{E}[W^2] \equiv \sigma^2 > 0$, *and* $\|\boldsymbol{w}(n)\|_4^4/m \to \mathbb{E}[W^4] < \infty$.

Further assume that each entry of $\boldsymbol{A}$ *satisfies the following regularity conditions:* $\mathbb{E}[A_{ij}] = 0$, $\mathbb{E}\left[\left(\sqrt{m}A_{ij}\right)^2\right] = 1$, *and* $\mathbb{E}\left[|\sqrt{m}A_{ij}|^p\right] = K_p < \infty$ *for some* $p > 4$, *where* K_p *is a constant independent of* n *(although we allow the distribution of* $\sqrt{m}A_{ij}$ *to be dependent on* n*).*

Fix $\lambda > 0$. *Then* $\mathrm{OPT}(\boldsymbol{A}) \to \mathrm{OPT}^*$ *in probability as* $n \to \infty$, *where* $\mathrm{OPT}^* = \mathrm{OPT}^*(\delta, \lambda, \sigma, p_{X_0})$ *a constant.*

Note that the case of Gaussian sensing matrix $A_{ij} \sim \mathsf{N}(0, 1/m)$ is covered by the above theorem. Also note that OPT^* is insensitive to the exact details of the distribution of the entries, and hence this is a universality phenomenon.

A closely related result was proven in [KM11], where universality was established for the cost of the box-constrained LASSO. The box constraint was crucial, and it is a non-trivial task to extend their proof to the LASSO being considered here. Some relevant universality results can be found further in [BLM15, OT15, MN17].

As in [MN17], the conditions on boundedness of $\|\boldsymbol{x}_0(n)\|_{10}^{10}/n$ and $\|\boldsymbol{w}(n)\|_4^4/m$ are not critical and can be weakened to boundedness of $\|\boldsymbol{x}_0(n)\|_{2+\epsilon}^{2+\epsilon}/n$ and $\|\boldsymbol{w}(n)\|_{2+\epsilon'}^{2+\epsilon'}/m$, for some $\epsilon, \epsilon' > 0$.

In the below, after a note on notations, we give a formula for OPT^* and state the analog of Theorem 1 for Gaussian sensing matrices. Then the main focus of the rest of this note is on proving Theorem 1, in which we first take a detour to Kashin's theorem in Section 2.

1.1 Notations

We use boldfaced lower-case letters (e.g. $\boldsymbol{x}$) for vectors and boldface upper-case letters (e.g. $\boldsymbol{A}$) for matrices. As usual, $\mathbb{N}$, $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{R}_{++}$ denote the set of natural numbers, real numbers, non-negative real numbers, and positive real numbers respectively. For $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, ..., n\}$. For a set $S \subseteq [n]$, $\bar{S}$ denotes its complement $[n] \setminus S$.

For $\boldsymbol{x} \in \mathbb{R}^n$ and a set $S \subseteq [n]$, $\boldsymbol{x}_S$ denotes a vector in $\mathbb{R}^n$ in which its i-th entry is equal to x_i if $i \in S$ and 0 otherwise. Likewise for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $\boldsymbol{A}_S$ denotes a matrix in $\mathbb{R}^{m\times n}$ in which its i-th column is equal to the i-th column of $\boldsymbol{A}$ if $i \in S$ and the all-zero column vector otherwise.

For $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we use $\sigma_{\min}(\boldsymbol{A})$ and $\sigma_{\max}(\boldsymbol{A})$ for respectively the smallest and largest singular values of $\boldsymbol{A}$. We follow the convention that singular values are non-negative. We denote the kernel of $\boldsymbol{A}$ and its orthogonal complement as $\ker(\boldsymbol{A})$ and $\ker(\boldsymbol{A})^\perp$.

1.2 Formula for OPT^* and the Gaussian Case

We give here the formula for OPT^*, which can be deduced using the approximate message passing algorithm. See [BM11, BM12] for more details. An alternative approach using the Gordon's

Gaussian min-max theorem, which is presented in [Sto13, OTH13, TOH15, TAH16], also yields a formula for OPT^*. In fact, the following result can be deduced from both of these lines of work.

Theorem 2. *Assume the setting in Theorem 1. Let $\boldsymbol{G}$ denote the Gaussian counterpart of $\boldsymbol{A}$, i.e. $G_{ij} \sim \mathsf{N}(0, 1/m)$ i.i.d. We have $\mathrm{OPT}(\boldsymbol{G}) \to \mathrm{OPT}^*$ in probability as $n \to \infty$.*

We denote by $\eta : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ the soft-thresholding function $\eta(x;\theta) = \mathrm{sign}(x)\max(0, |x| - \theta)$. Let $\tau_* > 0$ and $\alpha > 0$ be such that

$$\tau_*^2 = \sigma^2 + \frac{1}{\delta}\mathbb{E}\left[(\eta(X_0 + \tau_* Z; \alpha\tau_*) - X_0)^2\right], \tag{4}$$

$$\lambda = \alpha\tau_*\left[1 - \frac{1}{\delta}\mathbb{E}\left[\eta'(X_0 + \tau_* Z; \alpha\tau_*)\right]\right], \tag{5}$$

where $Z \sim \mathcal{N}(0,1)$ independent of X_0, and η' denotes the derivative of η w.r.t. the first argument. Then OPT^* is given by

$$\mathrm{OPT}^* = \frac{\delta\tau_*^2}{2}\left(1 - \frac{1}{\delta}\mathbb{P}(|X_0 + \tau_* Z| \geq \alpha\tau_*)\right)^2 + \lambda\mathbb{E}\left[\eta(|X_0 + \tau_* Z|; \alpha\tau_*)\right]. \tag{6}$$

2 Kashin's theorem

Kashin's theorem is a cornerstone result in asymptotic geometric analysis. It establishes the existence of a subspace of $\mathbb{R}^n$ of dimension proportional to n, in which the ℓ_2 and ℓ_1 norms are equivalent. Known constructions include the kernel of $\boldsymbol{A}$ or the range of $\boldsymbol{A}^T$, for $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ being a random sub-Gaussian matrix with $m < n$. See [MP03, LPR$^+$05].

The proof of Theorem 1 requires the following extension of Kashin's theorem, which we believe could be of independent interests.

Proposition 3. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be satisfying the regularity condition in Theorem 1, and $m/n \to \delta > 0$, $\delta \neq 1$. Then there exists a constant $c = c(\delta, K_4) \in (0,1)$ such that*

$$\mathbb{P}\left(\forall \boldsymbol{x} \in \ker(\boldsymbol{A}) :\ \|\boldsymbol{x}\|_1 \geq c\sqrt{n}\,\|\boldsymbol{x}\|_2\right) \to 1 \tag{7}$$

as $n \to \infty$, where $K_4 = \mathbb{E}\left[(\sqrt{m}A_{ij})^4\right]$. In particular, the above probability is lower-bounded by $1 - \exp(-\tilde{c}n)$ for some constant $\tilde{c} = \tilde{c}(\delta, K_4) > 0$.

With recent advances in random matrix theory, results of the above type could have been proven. We however are not aware of a reference, and hence provide here a proof.

To prove Proposition 3, we first rephrase a result from [Tik16], stated below.

Theorem 4. *For any $\zeta > 0$ and $\kappa \in (0,1)$, there exist $u, v > 0$ and $n_0 \in \mathbb{N}$ depending only on ζ and κ such that the following holds. Let $n_1, n_2 \in \mathbb{N}$ satisfy $n_1 \geq \max\{n_0, n_2/\kappa\}$, and $\boldsymbol{X} \in \mathbb{R}^{n_1\times n_2}$ with i.i.d. entries, such that for some $\omega > 0$,*

$$\sup_{\gamma\in\mathbb{R}} \mathbb{P}(|X_{ij} - \gamma| \leq \omega) \leq 1 - \zeta \tag{8}$$

Then $\mathbb{P}\left(\sigma_{\min}(\boldsymbol{X}) \leq \omega u\sqrt{n_1}\right) \leq \exp(-vn_1)$. In particular, the exponent constant on the right-hand side is

$$v = \min\left\{c_1(1-\kappa), c_2\left(\kappa^{1/2} - \kappa^{1/3}\right)^2, c_3\left(\kappa^{1/2} - \kappa\right)\left(\kappa^{1/4} - \kappa^{1/3}\right)\right\} \tag{9}$$

with positive constants c_1, c_2 and c_3 that depend only on ζ.

Corollary 5. *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be satisfying the regularity condition in Theorem 1, and $m = \lfloor \delta n \rfloor$ for $\delta \in (0,1)$. Let $K_4 = \mathbb{E}\left[\left(\sqrt{m}A_{ij}\right)^4\right]$. Fix $\kappa \in (0,\delta)$. There exist $u, v > 0$ that depend only on κ and K_4 such that*

$$\mathbb{P}\left(\forall S \subseteq [n] \;\; s.t. \;\; |S| = \kappa n : \; \sigma_{min}\left(\boldsymbol{A}_S\right) \leq u\right) \leq \exp\left[-\left(\delta v - \kappa \log\left(\frac{e}{\kappa}\right)\right) n\right] \tag{10}$$

for sufficiently large n. In particular, v is given by Eq. (9) with the constants c_1, c_2 and c_3 depending only on K_4.

Proof. Let $X_{ij} = \sqrt{m}A_{ij}$. One can easily show that

$$\sup_{\gamma\in\mathbb{R}} \mathbb{P}\left(|X_{ij} - \gamma| \leq \frac{1}{2}\right) \leq 1 - \min\left\{\frac{3}{11}, \frac{9}{|16\sqrt{K_4} - 7|}\right\} \tag{11}$$

Indeed, in particular, we can assume $\mathbb{E}\left[X_{ij}^3\right] \geq 0$, since the left-hand side is the same for X_{ij} and $-X_{ij}$. Then the Cauchy-Schwarz's inequality yields

$$\mathbb{P}\left((X_{ij}-\gamma)^2 \geq \frac{1}{4}\right) \geq \frac{\left(\mathbb{E}\left[(X_{ij}-\gamma)^2\right] - 1/4\right)^2}{\mathbb{E}\left[\left((X_{ij}-\gamma)^2 - 1/4\right)^2\right]} \geq \frac{\left(3/4 + \gamma^2\right)^2}{\sqrt{K_4} + 6\gamma^2 + \gamma^4} \tag{12}$$

which easily implies the claim. Hence, by Theorem 4, for sufficiently large n, there exist positive constants $u = u\left(\kappa, K_4\right)$ and $v = v\left(\kappa, K_4\right)$ such that

$$\mathbb{P}\left(\sigma_{min}\left(\boldsymbol{A}_S\right) \leq u\right) \leq \exp\left(-\delta v n\right) \tag{13}$$

for any $S \subseteq [n]$, $|S| = \kappa n$, and v is as described by the corollary statement. The union bound then yields

$$\mathbb{P}\left(\forall S \text{ s.t. } |S| = \kappa n : \; \sigma_{min}\left(\boldsymbol{A}_S\right) \leq u\right) \leq \binom{n}{\kappa n} \exp\left(-\delta v n\right) \leq \exp\left[-\left(\delta v - \kappa \log\left(\frac{e}{\kappa}\right)\right) n\right] \tag{14}$$

which completes the proof. □

Proof of Proposition 3. For $\delta > 1$, we have almost surely,

$$\lim_{n\to\infty} \sigma_{\min}\left(\boldsymbol{A}\right) = 1 - \frac{1}{\sqrt{\delta}} > 0 \tag{15}$$

by the Bai-Yin law, and therefore, $\ker\left(\boldsymbol{A}\right) = \{\boldsymbol{0}\}$ with probability converging to 1 as $n \to \infty$. Hence the statement is immediate in this case.

Consider $\delta \in (0,1)$. We prove the following equivalent statement: there exists a constant $c = c\left(\delta, K_4\right) \in (0,1)$ such that

$$\mathbb{P}\left(\forall \boldsymbol{x} \in \mathbb{R}^n \text{ s.t. } \|\boldsymbol{x}\|_1 < c\sqrt{n}\,\|\boldsymbol{x}\|_2 : \; \boldsymbol{x} \notin \ker\left(\boldsymbol{A}\right)\right) \to 1 \tag{16}$$

as $n \to \infty$. Consider $\boldsymbol{x} \neq \boldsymbol{0}$ such that $\|\boldsymbol{x}\|_1 \leq c\sqrt{n}\,\|\boldsymbol{x}\|_2$, for some $c \in (0,1)$. Let $S = \{i : \; |x_i| \geq \|\boldsymbol{x}\|_2 c/\left(\kappa\sqrt{n}\right)\}$ for some $\kappa \in (0,\delta)$. Then:

$$\|\boldsymbol{x}_{\bar{S}}\|_2^2 \leq \frac{c}{\kappa\sqrt{n}}\,\|\boldsymbol{x}\|_2\,\|\boldsymbol{x}_{\bar{S}}\|_1 \leq \frac{c}{\kappa\sqrt{n}}\,\|\boldsymbol{x}\|_2\,\|\boldsymbol{x}\|_1 \leq \frac{c^2}{\kappa}\,\|\boldsymbol{x}\|_2^2 \tag{17}$$

$$\|\boldsymbol{x}_S\|_2^2 = \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}_{\bar{S}}\|_2^2 \geq \|\boldsymbol{x}\|_2^2 \left(1 - \frac{c^2}{\kappa}\right) \tag{18}$$

$$|S| \leq \sum_{i \in S} \frac{|x_i|}{\|\boldsymbol{x}\|_2 c/\left(\kappa\sqrt{n}\right)} \leq \frac{\|\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_2 c/\left(\kappa\sqrt{n}\right)} \leq \kappa n \tag{19}$$

As $n \to \infty$, $\|\boldsymbol{A}\boldsymbol{x}_{\bar{S}}\|_2 \leq \left(1/\sqrt{\delta}+1\right)\|\boldsymbol{x}_{\bar{S}}\|_2$ eventually by the Bai-Yin law. Also, we have $\|\boldsymbol{A}\boldsymbol{x}_S\|_2 = \|\boldsymbol{A}_S\boldsymbol{x}_S\|_2 \geq \sigma_{min}\left(\boldsymbol{A}_S\right)\|\boldsymbol{x}_S\|_2$. By Corollary 5, for sufficiently large n, with probability at least $1 - \exp\left[-\left(\delta v - \kappa\log\left(e/\kappa\right)\right)n\right]$, $\sigma_{min}\left(\boldsymbol{A}_S\right) > u$ for some $u = u\left(\kappa, K_4\right) > 0$ and $v = v\left(\kappa, K_4\right) > 0$ given in Eq. (9). Then given $\delta v > \kappa\log\left(e/\kappa\right)$, with probability converging to 1 as $n \to \infty$,

$$\|\boldsymbol{A}\boldsymbol{x}\|_2 = \|\boldsymbol{A}\left(\boldsymbol{x}_S + \boldsymbol{x}_{\bar{S}}\right)\|_2 \geq \|\boldsymbol{A}\boldsymbol{x}_S\|_2 - \|\boldsymbol{A}\boldsymbol{x}_{\bar{S}}\|_2 \geq \left[u\sqrt{1 - \frac{c^2}{\kappa}} - \left(\frac{1}{\sqrt{\delta}}+1\right)\sqrt{\frac{c^2}{\kappa}}\right]\|\boldsymbol{x}\|_2 \tag{20}$$

which implies that any such $\boldsymbol{x}$ would not belong to $\ker\left(\boldsymbol{A}\right)$ if

$$u\sqrt{1 - \frac{c^2}{\kappa}} > \left(\frac{1}{\sqrt{\delta}}+1\right)\sqrt{\frac{c^2}{\kappa}}, \quad \delta v > \kappa\log\left(\frac{e}{\kappa}\right), \quad \frac{c^2}{\kappa} < 1, \quad 0 < \kappa < \delta, \quad 0 < c < 1.$$

From Eq. (9), we see that for a fixed $\delta \in (0,1)$, there exists $\kappa^* = \kappa^*\left(\delta, K_4\right) \in (0,\delta)$ such that the second constraint is satisfied with any $\kappa \in (0,\kappa^*)$. Then choosing

$$\kappa = \frac{\kappa^*}{2}, \quad c = \sqrt{\frac{u^2\kappa^*}{4\left[u^2 + \left(1 + 1/\sqrt{\delta}\right)^2\right]}} \tag{21}$$

completes the proof. □

3 Proof of Theorem 1

We first start with an application of Proposition 3 to the LASSO.

Lemma 6. *Consider the LASSO problem (2), with $\boldsymbol{A}$ satisfying the regularity condition of Theorem (1). Let $\hat{\boldsymbol{x}}\left(\boldsymbol{A}\right)$ be any of the LASSO minimizer. Then with probability converging to 1 as $n \to \infty$, $\|\hat{\boldsymbol{x}}\left(\boldsymbol{A}\right)\|_2^2/n \leq \mathsf{T} < \infty$ for some $\mathsf{T} = \mathsf{T}\left(\mathsf{M}_1, \mathsf{M}_2, K_4, \lambda, \sigma, \delta\right)$ a constant.*

Proof. Let $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}\left(\boldsymbol{A}\right)$ for brevity. Decompose $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}_{\|} + \hat{\boldsymbol{x}}_{\perp}$ where $\hat{\boldsymbol{x}}_{\|} \in \ker\left(\boldsymbol{A}\right)$ and $\hat{\boldsymbol{x}}_{\perp} \in \ker\left(\boldsymbol{A}\right)^{\perp}$. Let $c_0 = 100\left(\lambda\mathsf{M}_1 + \delta\sigma^2/2\right)$. We have

$$\frac{1}{2}\|\boldsymbol{A}\left(\hat{\boldsymbol{x}} - \boldsymbol{x}_0\right) - \boldsymbol{w}\|_2^2 + \lambda\|\hat{\boldsymbol{x}}\|_1 \leq C\left(\boldsymbol{x}_0, \boldsymbol{A}\right) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{x}_0\|_1 \leq nc_0 \tag{22}$$

for sufficiently large n. By Proposition 3, with probability converging to 1, for some constant $c = c\left(\delta, K_4\right) \in (0,1)$,

$$\left\|\hat{\boldsymbol{x}}_{\|}\right\|_2^2 \leq \frac{1}{c^2 n}\left\|\hat{\boldsymbol{x}}_{\|}\right\|_1^2 \overset{(a)}{\leq} \frac{1}{c^2 n}\left(\|\hat{\boldsymbol{x}}\|_1 + \|\hat{\boldsymbol{x}}_{\perp}\|_1\right)^2 \leq \frac{2}{c^2 n}\left(\|\hat{\boldsymbol{x}}\|_1^2 + \|\hat{\boldsymbol{x}}_{\perp}\|_1^2\right) \overset{(b)}{\leq} \frac{2}{c^2\lambda^2 n}n^2 c_0^2 + \frac{2}{c^2}\|\hat{\boldsymbol{x}}_{\perp}\|_2^2 \tag{23}$$

where (a) is by the triangular inequality and (b) is by Eq. (22). Also, with probability converging to 1,

$$\|\hat{\boldsymbol{x}}_\perp\|_2^2 \overset{(a)}{\leq} \left(\frac{\sqrt{\delta}}{1-\sqrt{\delta}}\right)^2 \|\boldsymbol{A}\hat{\boldsymbol{x}}_\perp\|_2^2 = \left(\frac{\sqrt{\delta}}{1-\sqrt{\delta}}\right)^2 \|\boldsymbol{A}\hat{\boldsymbol{x}}\|_2^2 \tag{24}$$

$$\overset{(b)}{\leq} \left(\frac{\sqrt{\delta}}{1-\sqrt{\delta}}\right)^2 \left(\|\boldsymbol{A}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}_0\right)-\boldsymbol{w}\|_2 + \|\boldsymbol{A}\boldsymbol{x}_0\|_2 + \|\boldsymbol{w}\|_2\right)^2 \tag{25}$$

$$\overset{(c)}{\leq} \left(\frac{\sqrt{\delta}}{1-\sqrt{\delta}}\right)^2 \left(\|\boldsymbol{A}\left(\hat{\boldsymbol{x}}-\boldsymbol{x}_0\right)-\boldsymbol{w}\|_2 + \left(\frac{1}{\sqrt{\delta}}+1\right)\|\boldsymbol{x}_0\|_2 + \|\boldsymbol{w}\|_2\right)^2 \tag{26}$$

$$\overset{(d)}{\leq} \left(\frac{\sqrt{\delta}}{1-\sqrt{\delta}}\right)^2 \left(\sqrt{2nc_0} + \left(\frac{1}{\sqrt{\delta}}+1\right)\sqrt{\mathsf{M}_2 n} + \sqrt{\delta\sigma^2 n}\right)^2 + 100n \tag{27}$$

where (a) and (c) are by the Bai-Yin law, (b) is by the triangular inequality, and (d) is by Eq. (22) and holds for n sufficiently large. The proof is complete by noticing that $\|\hat{\boldsymbol{x}}\|_2^2 = \left\|\hat{\boldsymbol{x}}_\parallel\right\|_2^2 + \|\hat{\boldsymbol{x}}_\perp\|_2^2$. □

We are ready to prove Theorem 1. The idea is to consider the elastic net:

$$\mathrm{OPT}_\rho\left(\boldsymbol{A}\right) = \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{n} C_\rho\left(\boldsymbol{x},\boldsymbol{A}\right), \quad C_\rho\left(\boldsymbol{x},\boldsymbol{A}\right) = \frac{1}{2}\left\|\boldsymbol{A}\left(\boldsymbol{x}-\boldsymbol{x}_0\right)-\boldsymbol{w}\right\|_2^2 + \lambda\|\boldsymbol{x}\|_1 + \frac{\rho}{2}\|\boldsymbol{x}\|_2^2 \tag{28}$$

for $\rho > 0$. We expect that as $\rho \to 0$, $\mathrm{OPT}_\rho\left(\boldsymbol{A}\right)$ yields $\mathrm{OPT}\left(\boldsymbol{A}\right)$. On the other hand, an inspection of [MN17] leads us to the following handy result. Let $h_k^- : \mathbb{R} \to [0,1]$ be a thrice-continuously differentiable and non-increasing mapping such that $h_k^-\left(x\right) = 1$ for $x \leq -1/k$ and $h_k^-\left(x\right) = 0$ for $x \geq 0$. Let $h_k^+\left(x\right) = h_k^-\left(x - 1/k\right)$. It is easy to see that $h_k^+\left(x\right) \to \mathbb{I}\left(x \leq 0\right)$ and $h_k^-\left(x\right) \to \mathbb{I}\left(x < 0\right)$ as $k \to \infty$, for any $x \in \mathbb{R}$. We have the following.

Theorem 7. *Assume the setting in Theorem 1. We have*

$$\mathbb{E}\left[h_k^-\left(\mathrm{OPT}_\rho\left(\boldsymbol{A}\right)-\ell\right)\right] - \mathbb{E}\left[h_k^-\left(\mathrm{OPT}_\rho\left(\boldsymbol{G}\right)-\ell\right)\right] \to 0 \tag{29}$$

as $n \to \infty$, for any $\ell \in \mathbb{R}$, any k and a given fixed $\rho > 0$.

Proof of Theorem 1. We have with probability converging to 1,

$$\mathrm{OPT}_\rho\left(\boldsymbol{A}\right) \geq \mathrm{OPT}\left(\boldsymbol{A}\right) \geq \mathrm{OPT}_\rho\left(\boldsymbol{A}\right) - \frac{\rho}{2n}\left\|\hat{\boldsymbol{x}}\left(\boldsymbol{A}\right)\right\|_2^2 \geq \mathrm{OPT}_\rho\left(\boldsymbol{A}\right) - \frac{\rho\mathsf{T}}{2} \tag{30}$$

where $\hat{\boldsymbol{x}}\left(\boldsymbol{A}\right)$ is any minimizer to $\mathrm{OPT}\left(\boldsymbol{A}\right)$, and the last inequality is by Lemma 6. Note that this also applies to $\boldsymbol{G}$. Then for any $\epsilon > 0$,

$$\mathbb{P}\left(\mathrm{OPT}\left(\boldsymbol{A}\right) \leq \mathrm{OPT}^* - \epsilon\right) \leq \mathbb{P}\left(\mathrm{OPT}_\rho\left(\boldsymbol{A}\right) \leq \mathrm{OPT}^* - \epsilon + \frac{\rho\mathsf{T}}{2}\right) + o_n\left(1\right) \tag{31}$$

$$\leq \mathbb{E}\left[h_k^+\left(\mathrm{OPT}_\rho\left(\boldsymbol{A}\right) - \mathrm{OPT}^* + \epsilon - \frac{\rho\mathsf{T}}{2}\right)\right] + o_n\left(1\right) \tag{32}$$

$$= \mathbb{E}\left[h_k^-\left(\mathrm{OPT}_\rho\left(\boldsymbol{A}\right) - \mathrm{OPT}^* + \epsilon - \frac{\rho\mathsf{T}}{2} - \frac{1}{k}\right)\right] + o_n\left(1\right) \tag{33}$$

$$\overset{(a)}{=} \mathbb{E}\left[h_k^-\left(\mathrm{OPT}_\rho\left(\boldsymbol{G}\right) - \mathrm{OPT}^* + \epsilon - \frac{\rho\mathsf{T}}{2} - \frac{1}{k}\right)\right] + o_n\left(1\right) \tag{34}$$

$$\leq \mathbb{P}\left(\mathrm{OPT}_\rho\left(\boldsymbol{G}\right) \leq \mathrm{OPT}^* - \epsilon + \frac{\rho \mathsf{T}}{2} + \frac{1}{k}\right) + o_n\left(1\right) \tag{35}$$

$$\leq \mathbb{P}\left(\mathrm{OPT}\left(\boldsymbol{G}\right) \leq \mathrm{OPT}^* - \epsilon + \frac{\rho \mathsf{T}}{2} + \frac{1}{k}\right) + o_n\left(1\right) \tag{36}$$

where we use Theorem 7 in step (a). For sufficiently small ρ and sufficiently large k, the right-hand-side of the above tends to 0 as $n \to \infty$, by Theorem 2. Likewise, $\mathbb{P}\left(\mathrm{OPT}\left(\boldsymbol{A}\right) \geq \mathrm{OPT}^* + \epsilon\right) \to 0$ as $n \to \infty$. The proof is complete. □

4 Discussion

The proof of Theorem 1 comprises of two main tools: Theorem 2 for the Gaussian case, and Theorem 7 for universality of the elastic net cost. Kashin's theorem plays a crucial role in establishing the bound (30) and hence uniform convergence of the elastic net cost to the LASSO cost when $n \to \infty$ and $\rho \to 0$.

We discuss why it is helpful to take a detour to the elastic net. The proof of Theorem 7 adopts the technique from [KM11], in particular, its proof for universality of the box-constrained LASSO cost. The box constraint here refers to the optimization domain being $\left\|\boldsymbol{x}\right\|_\infty \leq x_{\max}$ for some constant $x_{\max}$, instead of $\boldsymbol{x} \in \mathbb{R}^n$. It does not seem trivial to modify the technique when without this constraint. Although Lemma 6 implies that one can constrain the optimization domain of the LASSO to $\left\|\boldsymbol{x}\right\|_\infty \leq \sqrt{\mathsf{T}n}$, this bound does not appear sufficiently strong to prove universality. In [MN17], it is shown that one can restrict the optimization domain of the elastic net to $\left\|\boldsymbol{x}\right\|_\infty \leq g\left(n\right)$ for some $g\left(n\right) = O_\rho\left(n^{0.104}\right)$ (where the notation $O_\rho(\cdot)$ hides the explicit dependency on ρ). This is possible thanks to the fact $\rho > 0$, which makes the objective function strongly convex. This bound is sufficient to prove universality for the elastic net cost.

The bound is established by proving $\left\|\hat{\boldsymbol{x}}_\rho\right\|_\infty = O_\rho\left(n^{0.104}\right)$ with high probability, where $\hat{\boldsymbol{x}}_\rho$ is the elastic net minimizer. This may be much stronger than needed for dealing with the cost $\mathrm{OPT}\left(\boldsymbol{A}\right)$ or $\mathrm{OPT}_\rho\left(\boldsymbol{A}\right)$. In fact, it could be shown that, for

$$\mathrm{OPT}^B\left(\boldsymbol{A}\right) = \min_{\left\|\boldsymbol{x}\right\|_\infty \leq B} \frac{1}{n} C\left(\boldsymbol{x}, \boldsymbol{A}\right) \tag{37}$$

with $B > 0$ independent of n, we have

$$\lim_{B\to\infty} \lim_{n\to\infty} \mathbb{P}\left(\left|\mathrm{OPT}^B\left(\boldsymbol{A}\right) - \mathrm{OPT}\left(\boldsymbol{A}\right)\right| > \epsilon\right) = 0 \tag{38}$$

for any $\epsilon > 0$. This statement, along with universality of the box-constrained LASSO cost proven in [KM11], produces an alternative proof of Theorem 1. Our proof of this statement, however, still borrows the elastic net and certain techniques that have appeared in the proof of Theorem 7, so it is worth no more than this paragraph.

References

[BLM15] Mohsen Bayati, Marc Lelarge, and Andrea Montanari, *Universality in polytope phase transitions and message passing algorithms*, Ann. Appl. Probabil. **25** (2015), no. 2, 753–822.

[BM11] M. Bayati and A. Montanari, *The dynamics of message passing on dense graphs, with applications to compressed sensing*, IEEE Trans. Inf. Theory **57** (2011), no. 2, 764–785.

[BM12] ———, *The LASSO risk for gaussian matrices*, IEEE Trans. Inf. Theory **58** (2012), no. 4, 1997–2017.

[KM11] S. B. Korada and A. Montanari, *Applications of the Lindeberg principle in communications and statistical learning*, IEEE Trans. Inf. Theory **57** (2011), no. 4, 2440–2450.

[LPR+05] AE Litvak, A Pajor, M Rudelson, N Tomczak-Jaegermann, and R Vershynin, *Euclidean embeddings in spaces of finite volume ratio via random matrices*, Journal für die reine und angewandte Mathematik **2005** (2005), no. 589, 1–19.

[MN17] Andrea Montanari and Phan-Minh Nguyen, *Universality of the elastic net error*, Information Theory (ISIT), 2017 IEEE International Symposium on, IEEE, 2017, pp. 2338–2342.

[MP03] VD Milman and A Pajor, *Regularization of star bodies by random hyperplane cut off*, Studia Mathematica **159** (2003), 247–261.

[OT15] Samet Oymak and Joel A Tropp, *Universality laws for randomized dimension reduction, with applications*, arXiv preprint arXiv:1511.09433 (2015).

[OTH13] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi, *The squared-error of generalized lasso: A precise analysis*, arXiv preprint arXiv:1311.0830 (2013).

[Sto13] Mihailo Stojnic, *A framework to characterize performance of lasso algorithms*, arXiv preprint arXiv:1303.7291 (2013).

[TAH16] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, *Precise error analysis of regularized m-estimators in high-dimensions*, arXiv preprint arXiv:1601.06233 (2016).

[Tik16] Konstantin E. Tikhomirov, *The smallest singular value of random rectangular matrices with no moment assumptions on entries*, Israel Journal of Mathematics (2016), 1–26.

[TOH15] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi, *Regularized linear regression: A precise analysis of the estimation error*, Proc. 28th Conf. Learning Theory, 2015, pp. 1683–1709.