The LASSO risk: asymptotic results and real world examples

Authored by:

Andrea Montanari
Mohsen Bayati
Jos? Pereira

Abstract

We consider the problem of learning a coefficient vector x0 from noisy linear observation y=Ax0+w. In many contexts (ranging from model selection to image processing) it is desirable to construct a sparse estimator. In this case, a popular approach consists in solving an l1-penalized least squares problem known as the LASSO or BPDN. For sequences of matrices A of increasing dimensions, with iid gaussian entries, we prove that the normalized risk of the LASSO converges to a limit, and we obtain an explicit expression for this limit. Our result is the first rigorous derivation of an explicit formula for the asymptotic risk of the LASSO for random instances. The proof technique is based on the analysis of AMP, a recently developed efficient algorithm, that is inspired from graphical models ideas. Through simulations on real data matrices (gene expression data and hospital medical records) we observe that these results can be relevant in a broad array of practical applications.

1 Paper Body

Let x0 ? RN be an unknown vector, and assume that a vector y ? Rn of noisy linear measurements of x0 is available. The problem of reconstructing x0 from such measurements arises in a number of disciplines, ranging from statistical learning to signal processing. In many contexts the measurements are modeled by y = Ax0 + w ,

(1.1)

where A ? Rn?N is a known measurement matrix, and w is a noise vector. The LASSO or Basis Pursuit Denoising (BPDN) is a method for reconstructing the unknown vector x0 given y, A, and is particularly useful when one seeks sparse solutions. For given A, y, one considers the cost functions CA,y : RN ? R defined by 1 (1.2) CA,y (x) = ky ? Axk2 + ?kxk1 , 2 with ? ¿ 0. The original signal is estimated by x b(?; A, y) = argminx CA,y (x) .

(1.3)

In what follows we shall often omit the arguments A, y (and occasionally ?) from theP above notam tions. We will also use x b(?; N) to emphasize the N -dependence. Further kvkp ? (i=1 vip)1/p p denotes the ?p -norm of a vector v ? R (the subscript p will often be omitted if p = 2).

A large and rapidly growing literature is devoted to (i) Developing fast algorithms for solving the optimization problem (1.3); (ii) Characterizing the performances and optimality of the estimator x b. We refer to Section 1.3 for an unavoidably incomplete overview. 1

Despite such substantial effort, and many remarkable achievements, our understanding of (1.3) is not even comparable to the one we have of more classical topics in statistics and estimation theory. For instance, the best bound on the mean square error (MSE) of the estimator (1.3), i.e. on the quantity N ?1 kb x ? x0 k2 , was proved by Candes, Romberg and Tao [CRT06] (who in fact did not consider the LASSO but a related optimization problem). Their result estimates the mean square error only up to an unknown numerical multiplicative factor. Work by Candes and Tao [CT07] on the analogous Dantzig selector, upper bounds the mean square error up to a factor C log N , under somewhat different assumptions. The objective of this paper is to complement this type of ?rough but robust? bounds by proving asymptotically exact expressions for the mean square error. Our asymptotic result holds almost surely for sequences of random matrices A with fixed aspect ratio and independent gaussian entries. While this setting is admittedly specific, the careful study of such matrix ensembles has a long tradition both in statistics and communications theory and has spurred many insights [Joh06, Tel99]. Further, our main result provides asymptotically exact expressions for other operating characteristics of the LASSO as well (e.g., False Positive Rate and True positive Rate). We carried out simulations on real data matrices with continuous entries (gene expression data) and binary feature matrices (hospital medical records). The results appear to be quite encouraging. Although our rigorous results are asymptotic in the problem dimensions, numerical simulations have shown that they are accurate already on problems with a few hundreds of variables. Further, they seem to enjoy a remarkable universality property and to hold for a fairly broad family of matrices [DMM10]. Both these phenomena are analogous to ones in random matrix theory, where delicate asymptotic properties of gaussian ensembles were subsequently proved to hold for much broader classes of random matrices. Also, asymptotic statements in random matrix theory have been replaced over time by concrete probability bounds in finite dimensions. Of course the optimization problem (1.2) is not immediately related to spectral properties of the random matrix A. As a consequence, universality and non-asymptotic results in random matrix theory cannot be directly exported to the present problem. Nevertheless, we expect such developments to be foreseeable. Our proof is based on the analysis of an efficient iterative algorithm first proposed by [DMM09], and called AMP, for approximate message passing. The algorithm is inspired by belief-propagation on graphical models, although the resulting iteration is significantly simpler (and scales linearly in the number of nodes). Extensive simulations [DMM10] showed that, in a number of settings, AMP performances

are statistically indistinguishable to the ones of LASSO, while its complexity is essentially as low as the one of the simplest greedy algorithms. The proof technique just described is new. Earlier literature analyzes the convex optimization problem (1.3) ?or similar problems? by a clever construction of an approximate optimum, or of a dual witness. Such constructions are largely explicit. Here instead we prove an asymptotically exact characterization of a rather non-trivial iterative algorithm. The algorithm is then proved to converge to the exact optimum. Due to limited space in this paper we only state the main steps of the proof. More details are available in [BM10b] 1.1 Definitions In order to define the AMP algorithm, we denote by ? : R ? R+ ? R the soft thresholding function (x ? ? if x ¿ ?, 0 if ?? ? x ? ?, ?(x; ?) = (1.4) x + ? otherwise. The algorithm constructs a sequence of estimates xt ? RN , and residuals z t ? Rn , according to the iteration xt+1 = ?(A? z t + xt ; ?t), (1.5) kxt k0 t?1 z , n initialized with x0 = 0. Here A? denotes the transpose of matrix A, and kxt k0 is number of nonzero entries of xt . Given a scalar function f and a vector u ? Rm , we let f (u) denoteP the vector m (f (u1), . . . , f (um)) ? Rm obtained by applying f componentwise. Finally hui ? m?1 i=1 ui is m the average of the vector u ? R . z t = y ? Axt +

2

As already mentioned, we will consider sequences of instances of increasing sizes, along which the LASSO behavior has a non-trivial limit. Definition 1. The sequence of instances {x0 (N), w(N), A(N)}N ?N indexed by N is said to be a converging sequence if x0 (N) ? RN , w(N) ? Rn , A(N) ? Rn?N with n = n(N) such that n/N ? ? ? (0, ?), and in addition the following conditions hold: (a) The empirical1 distribution of the entries of x0 (N) converges weakly to a probability measure P 2 2 pX0 on R with bounded second moment. Further N ?1 N i=1 x0,i (N) ? EpX0 {X0 }. (b) The empirical distribution of the entries of w(N P) converges weakly to a probability measure pW on R with bounded second moment. Further n?1 ni=1 wi (N)2 ? EpW {W 2 }.

(c) If {ei }1?i?N , ei ? RN denotes the standard basis, then maxi?[N] kA(N)ei k2 , mini?[N] kA(N)ei k2 ? 1, as N ? ? where [N] ? {1, 2, . . . , N }.

For a converging sequence of instances, and an arbitrary sequence of thresholds {?t }t?0 (independent of N), the asymptotic behavior of the recursion (1.5) can be characterized as follows. Define the sequence {?t2 }t?0 by setting ?02 = ? 2 + E{X02 }/? (for X0 ? pX0 and ? 2 ? E{W 2 }, 2 W ? pW) and letting, for all t ? 0: ?t+1 = F(?t2 , ?t) with

1 E{ [?(X0 + ? Z; ?) ? X0]2 } , ? where Z ? N(0, 1) is independent of X0 . Notice that the function F depends on the law pX0 . F(? 2 , ?) ? ? 2 +

We say a function ? : R2 ? R is pseudo-Lipschitz if there exist a constant L ¿ 0 such that for all x, y ? R2 : —?(x) ? ?(y)— ? L(1 + kxk2 + kyk2)kx ? yk2 . (This is a special case of the definition used in [BM10a] where such a function is called pseudo-Lipschitz of order 2.) Our next proposition that was conjectured in [DMM09] and proved in [BM10a]. It shows that the behavior of AMP can be tracked by the above one dimensional recursion. We often refer to this prediction by state evolution. Theorem 1 ([BM10a]). Let {x0 (N), w(N), A(N)}N ?N be a converging sequence of instances with the entries of A(N) iid

normal with mean 0 and variance 1/n and let ? : R ? R ? R be a pseudoLipschitz function. Then, almost surely N n

o 1 X ? xt+1 , x = E ? ?(X + ? Z; ?), X , 0,i 0 t t 0 i N ?? N i=1

lim

(1.6)

where Z ? N(0, 1) is independent of X0 ? pX0 .

In order to establish the connection with the LASSO, a specific policy has to be chosen for the thresholds {?t }t?0 . Throughout this paper we will take ?t = ??t with ? is fixed. In other words, 2 the sequence {?t }t?0 is given by the recursion ?t+1 = F(?t2 , ??t). This choice enjoys several convenient properties [DMM09]. 1.2 Main result Before stating our results, we have to describe a calibration mapping between ? and ? that was introduced in [DMM10] (Propositions 2, 3 and Corollary 4). Their proofs are presented in [BM10b]. Let us start by stating some convenient properties of the state evolution recursion. Proposition 2 ([DMM09]). Let ?min = ?min (?) be the unique non-negative solution of the equation ? 2 (1 + ?2R)?(??) ? ??(?) = 2? , with ?(z) ? e?z /2 / 2? the standard gaussian density and z ?(z) ? ?? ?(x) dx.

For any ? 2 ¿ 0, ? ¿ ?min (?), the fixed point equation ? 2 = F(? 2 , ??) admits a unique solution. Denoting by ?? = ?? (?) this solution, we have limt?? ?t = ?? (?). Further the convergence takes

dF 2

place for any initial condition and is monotone. Finally d? 2 (? , ??) ¡ 1 at ? = ?? . 1

The probability distribution that puts a point mass 1/N at each of the N entries of the vector.

3

We then define the function ? 7? ?(?) on (?min (?), ?), by ?(?) ? ??? [1 ? 1? P{—X0 + ?? Z— ? ??? }].

This function defines a correspondence (calibration) between the sequence of thresholds {?t }t?0 and the regularization parameter ?. It should be intuitively clear that larger ? corresponds to larger thresholds and hence larger ? since both cases yield smaller estimates of x0 . In the following we willneed to invert this function. We thus define ? : (0, ?) ? (?min , ?) in such a way that ?(?) ? a ? (?min , ?) : ?(a) = ? .

The next result implies that the set on the right-hand side is non-empty and therefore the function ? 7? ?(?) is well defined. Proposition 3 ([DMM10]). The function ? 7? ?(?) is continuous on the interval (?min , ?) with ?(?min +) = ?? and lim??? ?(?) = ?.

Therefore the function ? 7? ?(?) satisfying ?(?) ? a ? (?min , ?) : ?(a) = ? exists. We will denote by A = ?((0, ?)) the image of the function ?. Notice that the definition of ? is a priori not unique. We will see that uniqueness follows from our main theorem. Examples of the mappings ? 2 7? F(? 2 , ??), ? 7? ?? (?) and ? 7? ?(?) are presented in [BM10b]. We can now state our main result. Theorem 2. Let {x0 (N), w(N), A(N)}N ?N be a converging sequence of instances with the entries of A(N) iid normal with mean 0 and variance 1/n. Denote by x b(?; N) the LASSO estimator for instance (x0 (N), w(N), A(N

)), with ? 2 , ? ¿ 0, P{X0 6= 0} and let ? : R ? R ? R be a pseudo-Lipschitz function. Then, almost surely N n

o 1 X ? x bi , x0,i = E ? ?(X0 + ?? Z; ??), X0 , N ?? N i=1

lim

(1.7)

where Z ? N(0, 1) is independent of X0 ? pX0 , ?? = ?? (?(?)) and ?? = ?(?)?? (?(?)). As a corollary, the function ? 7? ?(?) is indeed uniquely defined. Corollary 4. For any ?, ? 2 ¿ 0 there exists a unique ? ¿ ?min such that ?(?) = ? (with the function ? ? ?(?) defined by ?(?) = ??? [1 ? 1? P{—X0 + ?? Z— ? ??? }]. Hence the function ? 7? ?(?) is continuous non-decreasing with ?((0, ?)) ? A = (?0 , ?). The assumption of a converging problem-sequence is important for the result to hold, while the hypothesis of gaussian measurement matrices A(N) is necessary for the proof technique to be correct. On the other hand, the restrictions ?, ? 2 ¿ 0, and P{X0 6= 0} ¿ 0 (whence ?? 6= 0 using ?(?) = ??? [1 ? ?1 P{—X0 + ?? Z— ? ??? }]) are made in order to avoid technical complications due to degenerate cases. Such cases can be resolved by continuity arguments. 1.3 Related work The LASSO was introduced in [Tib96, CD95]. Several papers provide performance guarantees for the LASSO or similar convex optimization methods [CRT06, CT07], by proving upper bounds on the resulting mean square error. These works assume an appropriate ?isometry? condition to hold for A. While such condition hold with high probability for some random matrices, it is often difficult to verify them explicitly. Further, it is only applicable to very sparse vectors x0 . These restrictions are intrinsic to the worst-case point of view developed in [CRT06, CT07]. Guarantees have been proved for correct support recovery in [ZY06], under an appropriate ?irrepresentibility? assumption on A. While support recovery is an interesting conceptualization for some applications (e.g. model selection), the metric considered in the present paper (mean square error) provides complementary information and is quite standard in many different fields. Closer to the spirit of this paper [RFG09] derived expressions for the mean square error under the same model considered here. Similar results were presented recently in [KWT09, GBS09]. These papers argue that a sharp asymptotic characterization of the LASSO risk can provide valuable 4

guidance in practical applications. For instance, it can be used to evaluate competing optimization methods on large scale applications, or to tune the regularization parameter ?. Unfortunately, these results were non-rigorous and were obtained through the famously powerful ?replica method? from statistical physics [MM09]. Let us emphasize that the present paper offers two advantages over these recent developments: (i) It is completely rigorous, thus putting on a firmer basis this line of research; (ii) It is algorithmic in that the LASSO mean square error is shown to be equivalent to the one achieved by a low-complexity message passing algorithm.

2 Numerical illustrations Theorem 2 assumes that the entries of matrix A are iid gaussians. We expect however that our predictions to be robust and hold for much larger family of matrices. Rigorous evidence in this direction is presented in [KM10] where the normalized cost C(b x)/N is shown to have a limit as N ?

? which is universal with respect to random matrices A with iid entries. (More precisely, it is universal if E{Aij } = 0, E{A2ij } = 1/n and E{A6ij } ? C/n3 for a uniform constant C.) Further, our result is asymptotic, while and one might wonder how accurate it is for instances of moderate dimensions. Numerical simulations were carried out in [DMM10] and suggest that the result is robust and relevant already for N of the order of a few hundreds. As an illustration, we present in Figures 1-3 the outcome of such simulations for four types of real data and random matrices. We generated the signal vector randomly with entries in {+1, 0, ?1} and P(x0,i = +1) = P(x0,i = ?1) = 0.05. The noise vector w was generated by using i.i.d. N(0, 0.2) entries. We obtained the optimum estimator x b using OWLQN and l1 ls, packages for solving large-scale l1-regularized regressions [KKL+ 07], [AJ07]. We used 40 values of ? between .05 and 2 and N equal to 500, 1000, and 2000. For each case, the point (?, MSE) was plotted and the results are shown in the figures. Continuous lines corresponds to the asymptotic prediction by Theorem 2 for

2 ?(a, b) = (a ? b)2 , namely MSE = limN ?? N ?1 kb x ? x0 k2 = E ?(X0 + ?? Z; ??) ? X0 = ?(??2 ? ? 2).

The agreement is remarkably good already for N, n of the order of a few hundreds, and deviations are consistent with statistical fluctuations. The four figures correspond to measurement matrices A:

Figure 1(a): Data consist of 2253 measurements of expression level of 7077 genes (this data is provided to us by Broad Institute). From this matrix we took sub-matrices A of aspect ratio ? for each N . The entries were continuous variables. We standardized all columns of A to have mean 0 and variance 1. Figure 1(b): From a data set of 1932 patient records we extracted 4833 binary features describing demographic information, medical history, lab results, medications etc. The 0-1 matrix was sparse (with only 3.1% non-zero entries). Similar to genes data, for each N , the sub-matrices A with aspect ratio ? were selected and standardized. ? Figure 2(a): ? Random ?1 matrices with aspect ratio ?. Each entry is independently equal to +1/ n or ?1/ n with equal probability.

Figure 2(b): Random gaussian matrices with aspect ratio ? and iid N(0, 1/n) entries (as in Theorem 2). Notice the behavior appears to be essentially indistinguishable. Also the asymptotic prediction has a minimum as a function of ?. The location of this minimum can be used to select the regularization parameter. Further empirical analysis is presented in [BBM10]. For the second data set ?patient records? we repeated the simulation 20 times (each time with fresh x0 and w) and obtained the average and standard error for MSE, False Positive Rate (FPR) and True Positive Rate (TPR). The results with error bars are shown in Figure 3. The length of each error bar 5

(a) Gene expression data

(b) Hospital records

0.4

0.4 N=500 N=1000 N=2000 Prediction

0.3

0.3

0.25
0.25
0.2
0.2
0.15
0.15
0.1
0.1
0.05 0
0.5
1
?
1.5
2
N=500 N=1000 N=2000 Prediction
0.35
MSE
MSE
0.35
0.05 0
2.5
0.5
1
?
1.5
2
2.5

Figure 1: Mean square error (MSE) as a function of the regularization parameter ? compared to the asymptotic prediction for ? = .5 and ? 2 = .2. In plot (a) the measurement matrix A is a real valued (standardized) matrix of gene expression data and in plot (b) A is a (standardized) 0-1 feature matrix of hospital records. Each point in these plots is generated by finding the LASSO predictor x b using a measurement vector y = Ax0 + w for an independent signal vector x0 and an independent noise vector w. (a) ?1 matrices

(b) Gaussian matrices

0.4
0.4 N=500 N=1000 N=2000 Prediction
0.3
0.3
0.25
0.25
0.2
0.2
0.15
0.15
0.1

0.1
0.05 0
0.5
1
?
1.5
2
N=500 N=1000 N=2000 Prediction
0.35
MSE
MSE
0.35
0.05 0
2.5
0.5
1
?
1.5
2
2.5

? Figure 2: As in Figure 1, but the measurement matrix A has iid entries that are equal to ?1/ n with equal probabilities in plot (a), and has iid N(0, 1/n) entries in plot (b). Additionally, each point in these plots uses an independent matrix A. is equal to twice the standard error (in each direction). FPR and TPR are calculated using PN PN I{?xi 6=0} I{xi,0 6=0} xi 6=0} I{xi,0 =0} i=1 I{? FPR ? , TPR ? i=1 , PN PN i=1 I{xi,0 =0} i=1 I{xi,0 6=0}

(2.1)

where I{S} = 1 if statement S holds and I{S} = 0 otherwise. The predictions for FPR and TPR are obtained by applying Theorem 2 to ?f pr (a, b) ? I{a6=0} I{b=0} and ?tpr (a, b) = I{a6=0} I{b6=0} , which yields 1 1 lim FPR = 2?(??), lim TPR = ?(?? +) + ?(?? ?) (2.2) N ?? N ?? ?? ?? where ? is defined in Proposition 2. Note that functions ?f pr (a, b) and ?tpr (a, b) are not pseudoLipschitz but the limits (2.2) follow from Theorem 2 via standard weak-convergence arguments.

3 A structural property and proof of the main theorem We will prove the following theorem which implies our main result, Theorem 2. Theorem 3. Assume the hypotheses of Theorem 2. Denote by {xt (N)}t?0 the sequence of estimates produced by AMP. Then limt?? limN ?? N ?1 kxt (N) ? x b(?; N)k22 = 0, almost surely. 6

0.4
0.5 N500 N1000 N2000 Prediction
0.35
0.4 False Positive Rate
0.3
MSE
N=500 N=1000 N=2000 Prediction

0.45
0.25 0.2
0.35 0.3 0.25 0.2 0.15
0.15
0.1 0.1 0.05 0.05 0
0.5
1
?
1.5
2
0 0
2.5
0.5
1
?
1.5
2
2.5
0.8 N=500 N=1000 N=2000 Prediction
0.7
True Positive Rate
0.6 0.5 0.4 0.3 0.2 0.1 0 0
0.5
1
?
1.5
2
2.5

Figure 3: Average of MSE, FPR and TPR versus ? for medical data, using 20 samples per ? and N . All parameters are similar to Figure 1(b). Error bars are twice the standard errors (in each direction). The rest of the paper is devoted to the proof of this theorem. Section 3.1 proves a structural property that is the key tool in this proof. Section 3.2 uses this property together with a few lemmas to prove Theorem 3. Proofs of lemmas and more details can be found in [BM10b]. The proof of Theorem 2 follows immediately. Since when ? is Lipschitz there is a constant B where PN PN — N1 i=1 ?(xt+1 , x0,i) ? N1 xi , x0,i)— ? Bkxt+1 ? x bk2 . We then obtain i i=1 ?(b N N

1 X 1 X ?(b xi , x0,i) = lim lim ? xt+1 , x0,i = E{?(?(X0 + ?? Z; ??), X0)} , i t?? N ?? N N ?? N i=1 i=1

lim

where we used Theorem 1 and Proposition 2. The case of pseudo-Lipschitz ? is a straightforward generalization. Some notations. For any non-empty subset S of [m] and any k ? m matrix M we refer by MS to the k by —S— sub-matrix of M thatPcontains only the columns of M corresponding to S. Also define m 1 m the scalar product hu, vi ? m i=1 ui vi for u, v ? R . Finally, the subgradient of a convex function f : Rm ? R at point x ? Rm is denoted by ?f (x). In

particular, remember that the subgradient of the ?1 norm, x 7? kxk1 is given by

?kxk1 = v ? Rm such that —vi — ? 1 ?i and xi 6= 0 ? vi = sign(xi) . (3.1) 3.1 A structural property of the LASSO cost function

One main challenge in the proof of Theorem 2 lies in the fact that the function x 7? CA,y (x) is not ?in general? strictly convex. Hence there can be, in principle, vectors x of cost very close to the optimum and nevertheless far from the optimum. The following Lemma provides conditions under which this does not happen. Lemma 1. There exists a function ?(?, c1 , . . . , c5) such that the following happens. If x, r ? RN satisfy the following conditions: 7

? (1) krk2 ? c1 N ;

(2) C(x + r) ? C(x);

(3) There exists a subgradient sg(C, x) ? ?C(x) with ksg(C, x)k2 ?

? N ?;

(4) Let v ? (1/?)[A? (y ? Ax) + sg(C, x)] ? ?kxk1 , and S(c2) ? {i ? [N] : —vi — ? 1 ? c2 }. Then, for any S ? ? [N], —S ? — ? c3 N , we have ?min (AS(c2)?S ?) ? c4

(5) The maximum and minimum non-zero singular value of A satisfy c?1 ? ?min (A)2 ? 5 2 ?max (A) ? c5 . ? Then krk2 ? N ?(?, c1 , . . . , c5). Further for any c1 , . . . , c5 ¿ 0, ?(?, c1 , . . . , c5) ? 0 as ? ? 0.

Further, if ker(A) = {0}, the same conclusion holds under conditions 1, 2, 3, and 5. 3.2 Proof of Theorem 3

The proof is based on a series of Lemmas that are used to check the assumptions of Lemma 1 The next lemma implies that submatrices of A constructed using the first t iterations of the AMP algorithm are non-singular (more precisely, have singular values bounded away from 0). Lemma 2. Let S ? [N] be measurable on the ?-algebra St generated by {z 0 , . . . , z t?1 } and {x0 + A? z 0 , . . . , xt?1 + A? z t?1 } and assume —S— ? N (? ? c) for some c ¿ 0. Then there exists a1 = a1 (c) ¿ 0 (independent of t) and a2 = a2 (c, t) ¿ 0 (depending on t and c) such that minS ? {?min (AS?S ?) : S ? ? [N], —S ? — ? a1 N } ? a2 , with probability converging to 1 as N ? ?. We will apply this lemma to a specific choice of the set S. Namely, defining vt ?

1 ?t?1

(xt?1 + A? z t?1 ? xt) ,

(3.2)

our last lemma shows convergence of a particular sequence of sets provided by v t .

Lemma 3. Fix ? ? (0, 1) and let the sequence {St (?)}t?0 be defined by St (?) ? i ? [N] :

—vit — ? 1 ?? . For any ? ¿ 0 there exists t? = t? (?, ?) ¡ ? such that, for all t2 ? t1 ? t? : limN ?? P —St2 (?) St1 (?)— ? N ? = 0. The last two lemmas imply the following.

Proposition 5. There exist constants ?1 ? (0, 1), ?2 , ?3 ¿ 0 and tmin ¡ ? such that, for any t ? tmin , min ?min (ASt (?1)?S ?) : S ? ? [N] , —S ? — ? ?2 N ? ?3 with probability converging to 1 as N ? ?. Proof of Theorem 3. We apply Lemma 1 to x = xt , the AMP estimate and r = x b ? xt the

distance from the LASSO optimum. The thesis follows by checking conditions 1?5. Namely we need to show that there exists constants c1 , . . . , c5 ¿ 0 and, for each ? ¿ 0 some t = t(?) such that 1?5 hold with probability going to 1 as N ? ?. Condition 1 holds since limN ?? hb x, x bi and limN ?? hxt , xt i for all t are finite. Condition 2 is immediate since x + r = x b minimizes C(?).

Conditions 3-4. Take v = v t as defined in Eq. (3.2). Using the definition (1.5), it is easy to check that v t ? ?kxk1 . Further it can be shown that v t = (1/?)[A? (y ? Axt) + sg(C, xt)], with sg(C, xt) a subgradient satisfying limt?? limN ?? N ?1 ksg(C, xt)k2 = 0. This proves condition 3 and condition 4 holds by Proposition 5. Condition 5 follows from standard limit theorems on the singular values of Wishart matrices.

Acknowledgement This work was partially supported by a Terman fellowship, the NSF CAREER award CCF-0743978, the NSF grant DMS-0806211 and a Portuguese Doctoral FCT fellowship.

8

2 References

[AJ07]

G. Andrew and G. Jianfeng, Scalable training of l1 -regularized log-linear models, Proceedings of the 24th international conference on Machine learning, 2007, pp. 33?40.

[BBM10] M. Bayati, J .A. Bento, and A. Montanari, The LASSO risk: asymptotic results and real world examples, Long version (in preparation), 2010. [BM10a]

[BM10b]

M. Bayati and A. Montanari, The dynamics of message passing on dense graphs, with applications to compressed sensing, Proceedings of IEEE International Symposium on Inform. Theory (ISIT), 2010, Longer version in http://arxiv.org/abs/1001.3448. , The LASSO risk for gaussian matrices, 2010, preprint available in http://arxiv.org/abs/1008.2581.

[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.

[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.

[CT07]

E. Candes and T. Tao, The Dantzig selector: statistical estimation when p is much larger than n, Annals of Statistics 35 (2007), 2313?2351.

[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, Preprint, 2010. [GBS09]

D. Guo, D. Baron, and S. Shamai, A single-letter characterization of optimal noisy compressed sensing, 47th Annual Allerton Conference (Monticello, IL), September 2009.

[Joh06]

I. Johnstone, High Dimensional Statistical Inference and Random Matrices, Proc. International Congress of Mathematicians (Madrid), 2006.

[KKL+ 07] S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, A method for large-scale ?1 -regularized least squares., IEEE Journal on Selected Topics in Signal Processing 1 (2007), 606?617. [KM10]

S. Korada and A. Montanari, Applications of Lindeberg Principle in Communications and Statistical Learning, preprint available in http://arxiv.org/abs/1004.0557, 2010.

[KWT09] Y. Kabashima, T. Wadayama, and T. Tanaka, A typical reconstruction limit for compressed sensing based on lp-norm minimization, J.Stat. Mech. (2009), L09003. [MM09] M. M?ezard and A. Montanari, Information, Physics and Computation, Oxford University Press, Oxford, 2009. [RFG09] S. Rangan, A. K. Fletcher, and V. K. Goyal, Asymptotic analysis of map estimation via the replica method and applications to compressed sensing, PUT NIPS REF, 2009. [Tel99] E. Telatar, Capacity of Multi-antenna Gaussian Channels, European Transactions on Telecommunications 10 (1999), 585?595. [Tib96] R. Tibshirani, Regression shrinkage and selection with the lasso, J. Royal. Statist. Soc B 58 (1996), 267?288. [ZY06]

P. Zhao and B. Yu, On model selection consistency of Lasso, The Journal of Machine Learning Research 7 (2006), 2541?2563.

9