More Supervision, Less Computation: Statistical-Computational Tradeoffs in Weakly Supervised Learning

Authored by:

Han Liu
Constantine Caramanis
Zhaoran Wang
Zhuoran Yang
Xinyang Yi

Abstract

We consider the weakly supervised binary classification problem where the labels are randomly flipped with probability $1-alpha$. Although there exist numerous algorithms for this problem, it remains theoretically unexplored how the statistical accuracies and computational efficiency of these algorithms depend on the degree of supervision, which is quantified by $alpha$. In this paper, we characterize the effect of $alpha$ by establishing the information-theoretic and computational boundaries, namely, the minimax-optimal statistical accuracy that can be achieved by all algorithms, and polynomial-time algorithms under an oracle computational model. For small $alpha$, our result shows a gap between these two boundaries, which represents the computational price of achieving the information-theoretic boundary due to the lack of supervision. Interestingly, we also show that this gap narrows as $alpha$ increases. In other words, having more supervision, i.e., more correct labels, not only improves the optimal statistical accuracy as expected, but also enhances the computational efficiency for achieving such accuracy.

1 Paper Body

Practical classi?cation problems usually involve corrupted labels. Speci?cally, let {(xi , zi)}ni=1 be n independent data points, where xi ? Rd is the covariate vector and zi ? {0, 1} is the uncorrupted label. Instead of observing {(xi , zi)}ni=1 , we observe {(xi , yi)}ni=1 in which yi is the corrupted label. In detail, with probability (1 ? ?), yi is chosen uniformly at random over {0, 1}, and with probability ?, yi = zi . Here ? ? [0, 1] quanti?es the degree of

supervision: a larger ? indicates more supervision since we have more uncorrupted labels in this case. In this paper, we are particularly interested in the effect of ? on the statistical accuracy and computational ef?ciency for parameter estimation in this problem, particularly in the high dimensional settings where the dimension d is much larger than the sample size n. There exists a vast body of literature on binary classi?cation problems with corrupted labels. In particular, the study of randomly perturbed labels dates back to [1] in the context of random classi?cation noise model. See, e.g., [12, 20] for a survey. Also, classi?cation problems with missing labels are also extensively studied in the context of semi-supervised or weakly supervised learning by [14, 17, 21], among others. Despite the extensive study on this problem, its information-theoretic and computational boundaries remain unexplored in terms of theory. In a nutshell, the informationtheoretic boundary refers to the optimal statistical accuracy achievable by any algorithms, while the computational boundary refers to the optimal statistical accuracy achievable by the algorithms under a computational budget that has a polynomial dependence on the problem scale (d, n). Moreover, it remains unclear how these two boundaries vary along with ?. One interesting question to ask is 29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

how the degree of supervision affects the fundamental statistical and computational dif?culties of this problem, especially in the high dimensional regime. In this paper, we sharply characterize both the information-theoretic and computational boundaries of the weakly supervised binary classi?cation problems under the minimax framework. Speci?cally, we consider the Gaussian generative model where X—Z = z ? N (?z , ?) and z ? {0, 1} is the true label. Suppose {(xi , zi)}ni=1 are n independent samples of (X, Z). We assume that {yi }ni=1 are generated from {zi }ni=1 in the aforementioned manner. We focus on the high dimensional regime, where d ? n and ?1 ? ?0 is s-sparse, i.e., ?1 ? ?0 has s nonzero entires. We are interested in estimating ?1 ? ?0 from the observed samples {(xi , yi)}ni=1 . By a standard reduction argument [24], the fundamental limits of this estimation task are captured by a hypothesis testing problem, namely, H0 : ?1 ? ?0 = 0 versus H1 : ?1 ? ?0 is s-sparse and (?1 ? ?0)? ??1 (?1 ? ?0) := ?n ¿ 0,

(1.1)

where ?n denotes the signal strength that scales with n. Consequently, we focus on studying the fundamental limits of ?n for solving this hypothesis testing problem. ?n ??

E?cient

? s2 s log d ?n = o ? 2 , n ? n ?? ? s log d s log d ?n = ? ? 2 n ? n

?n = ?

??

s2 s log d ? 2 n ? n

?

Intractable Impossible 0

?

?n = o

??
s log d s log d ? 2 n ? n
?
1

Figure 1: Computational-statistical phase transitions for weakly supervised binary classi?cation. Here ? denotes the degree of supervision, i.e., the label is corrupted to be uniformly random with probability 1 ? ?, and ?n is the signal strength, which is de?ned in (1.1). Here a ? b denotes min{a, b}.

Our main results are illustrated in Figure 1. Speci?cally, we identify the impossible, intractable, and ef?cient regimes for the statistical-computational phase transitions under certain regularity conditions. ? (i) For ?n = o[s log d/n ? (1/?2 ? s log d/n)], any algorithm is asymptotically powerless in solving the hypothesis ? testing problem. ? (ii) For ?n = ?[s log d/n ? (1/?2 ? s log d/n)] and ?n = o[s2 /n ? (1/?2 ? s log d/n)], any tractable algorithm that has a polynomial oracle complexity under an extension of the statistical query model [18] is asymptotically powerless. We will rigorously de?ne the computational model in ?2. ? (iii) For ?n = ?[s2 /n ? (1/?2 ? s log d/n)], there is an ef?cient algorithm with a polynomial oracle complexity that is asymptotically powerful in solving the testing problem. ? ? Here s log d/n ? (1/?2 ? s log d/n) gives the information-theoretic boundary, while s2 /n ? (1/?2 ? s log d/n) gives the computational boundary. Moreover, by a reduction from the estimation problem to the testing problem, these boundaries for testing imply the ones for estimating ?2 ? ?1 as well. Consequently, there exists a signi?cant gap between the computational and information-theoretic boundaries for small ?. In other word, to achieve the information-theoretic boundary, one has to pay the price of intractable computation. As ? tends to one, this gap between computational and information-theoretic boundaries narrows and eventually vanishes. This indicates that, having more supervision not only improves the statistical accuracy, as shown by the decay of information-theoretic boundary in Figure 1, but more importantly, enhances the computational ef?ciency by reducing the computational price for attaining information-theoretic optimality. This phenomenon ? ?more supervision, less computation? ? is observed for the ?rst time in this paper. 1.1

More Related Work, Our Contribution, and Notation

Besides the aforementioned literature on weakly supervised learning and label corruption, our work is also connected to a recent line of work on statistical-computational tradeoffs [2?5, 8, 13, 15, 19, 26?28]. In comparison, we quantify the statistical-computational tradeoffs for weakly supervised learning for the ?rst time. Furthermore, our results are built on an oracle computational model 2

in [8] that slightly extends the statistical query model [18], and hence do not hinge on unproven conjectures on computational hardness like planted clique. Compared with our work, [8] focuses on the computational hardness of learning heterogeneous models, whereas we consider the interplay between supervision and statistical-computational tradeoffs. A similar computational model is used in [27] to study structural normal mean model and principal component analysis,

which exhibit different statistical-computational phase transitions. In addition, our work is related to sparse linear discriminant analysis and two-sample testing of sparse means, which correspond to our special cases of ? = 1 and ? = 0, respectively. See, e.g., [7, 23] for details. In contrast with their results, our results capture the effects of ? on statistical and computational tradeoffs. In summary, the contribution of our work is two-fold: (i) We characterize the computational and statistical boundaries of the weakly supervised binary classi?cation problem for the ?rst time. Compared with existing results for other models, our results do not rely on unproven conjectures. (ii) Based on our theoretical characterization, we propose the ?more supervision, less computation? phenomenon, which is observed for the ?rst time. Notation. We denote the ?2 -divergence between two distributions P, Q by D?2 (P, Q). For two nonnegative sequences an , bn indexed by n, we use an = o(bn) as a shorthand for limn?? an /bn = 0. We say an = ?(bn) if an /bn ? c for some absolute constant c ¿ 0 when n is suf?ciently large. We use a ? b and a ? b to denote max{a, b} and min{a, b}, respectively. For any positive integer k, we denote {1, 2, . . . , k} by [k]. For v ? Rd , we denote by ?v?p the ?p -norm of v. In addition, we denote the operator norm of a matrix A by ———A———2 .

2

Background

In this section, we formally de?ne the statistical model for weakly supervised binary classi?cation. Then we follow it with the statistical query model that connects computational complexity and statistical optimality. 2.1

Problem Setup

Consider the following Gaussian generative model for binary classi?cation. For a random vector X ? Rd and a binary random variable Z ? {0, 1}, we assume X—Z = 0 ? N (?0 , ?),

X—Z = 1 ? N (?1 , ?),

(2.1)

where P(Z = 0) = P(Z = 1) = 1/2. Under this model, the optimal classi?er by Bayes rule corresponds to the Fisher?s linear discriminative analysis (LDA) classi?er. In this paper, we focus on the noisy label setting where true label Z is replaced by a uniformly random label in {0, 1} with probability 1 ? ?. Hence, ? characterizes the degree of supervision in the model. In speci?c, if ? = 0, we observe the true label Z, thus the problem belongs to supervised learning. Whereas if ? = 1, the observed label is completely random, which contains no information of the model in (2.1). This setting is thus equivalent to learning a Gaussian mixture model, which is an unsupervised problem. In the general setting with noisy labels, we denote the observed label by Y , which is linked to the true label Z via P(Y = Z) = (1 + ?)/2, P(Y = 1 ? Z) = (1 ? ?)/2. (2.2)

We consider the hypothesis testing problem of detecting whether ?0 ?= ?1 given n i.i.d. samples {yi , xi }ni=1 of (Y, X), namely H0 : ?0 = ?1 versus H1 : ?0 ?= ?1 .

(2.3)

We focus on the high dimensional and sparse regime, where d ? n and ?0 ? ?1 is s-sparse, i.e., ?0 ? ?1 ? B0 (s), where B0 (s) := {? ? Rd : ???0 ?

s}. Throughout this paper, use the sample size n to drive the asymptotics. We introduce a shorthand notation ? := (?0 , ?1 , ?, ?) to represent the parameters of the aforementioned model. Let P? be the joint distribution of (Y, X) under our statistical model with parameter ?, and Pn? be the product distribution of n i.i.d. samples accordingly. We denote the parameter spaces of the null and alternative hypotheses by G0 and G1 respectively. For any test function ? : {(yi , xi)}ni=1 ? {0, 1}, the classical testing risk is de?ned as the summation of 3

type-I and type-II errors, namely Rn (?; G0 , G1) := sup Pn? (? = 1) + sup Pn? (? = 0). ??G0

??G1

The minimax risk is de?ned as the smallest testing risk of all possible test functions, that is, Rn? (G0 , G1) := inf Rn (?; G0 , G1), ?

(2.4)

where the in?mum is taken over all measurable test functions. Intuitively, the separation between two Gaussian components under H1 and the covariance matrix ? together determine the hardness of detection. To characterize such dependence, we de?ne the signalto-noise ratio (SNR) as ?(?) := (?0 ? ?1)? ??1 (?0 ? ?1). For any nonnegative sequence {?n }n?1 , let G1 (?n) := {? : ?(?) ? ?n } be a sequence of alternative parameter spaces with minimum separation ?n . The following minimax rate characterizes the information-theoretic limits of the detection problem. De?nition 2.1 (Minimax rate). We say a sequence {?n? }n?1 is a minimax rate if ? For any sequence {?n }n?1 satisfying ?n = o(?n?), we have limn?? Rn? [G0 , G1 (?n)] = 1; ? For any sequence {?n }n?1 satisfying ?n = ?(?n?), we have limn?? Rn? [G0 , G1 (?n)] = 0. The minimax rate in De?nition 2.1 characterizes the statistical dif?culty of the testing problem. However, it fails to shed light on the computational ef?ciency of possible testing algorithms. The reason is that this concept does not make any computational restriction on the test functions. The minimax risk in (2.4) might be attained only by test functions that have exponential computational complexities. This limitation of De?nition 2.1 motivates us to study statistical limits under computational constraints. 2.2

Computational Model

Statistical query models [8?11, 18, 27] capture computational complexity by characterizing the total number of rounds an algorithm interacts with data. In this paper, we consider the following statistical query model, which admits bounded query functions but allows the responses of query functions to be unbounded. De?nition 2.2 (Statistical query model). In the statistical query model, an algorithm A is allowed to query an oracle T rounds, but not to access data {(yi , xi)}ni=1 directly. At each round, A queries the oracle r with a query function q ? QA , in which QA ? {q : {0, 1} ? Rd ? [?M, M]} denotes the query space of A . The oracle r outputs a realization of a random variable Zq ? R satisfying ? ? ? ? ? P —Zq ? E[q(Y, X)]— ? ?q ? 1 ? 2?, where q?QA

?q = [?(QA) + log(1/?)] ? M/n

?? ? 2[?(QA) + log(1/?)] ? (M 2 ? {E[q(Y, X)]}2) n. (2.5)

Here ?q ¿ 0 is the tolerance parameter and ? ? [0, 1) is the tail probability.

The quantity ?(QA) ? 0 in ?q measures the capacity of QA in logarithmic scale, e.g., for countable QA , ?(QA) = log(—QA —). The number T is de?ned as the oracle complexity. We denote by R[?, n, T, ?(QA)] the set of oracles satisfying (2.5), and by A(T) the family of algorithms that queries an oracle no more than T rounds. This version of statistical query model is used in [8], and reduces to the VSTAT model proposed in [9?11] by the transformation q?(y, x) = q(y, x)/(2M) + 1/2 for any q ? QA . The computational model in De?nition 2.2 enables us to handle query functions that are bounded by an unknown and ?xed number M . Note that that by incorporating the tail probability ?, the response Zq is allowed to be unbounded. To understand the intuition behind De?nition 2.2, we remark that (2.5) resembles the Bernstein?s inequality for bounded random variables [25] ? ?? ? ? ? ? ?1 n ? t2 q(Yi , Xi) ? E[q(Y, X)]?? ? t ? 2 exp . (2.6) P ?? n 2Var[q(Y, X)] + M t i=1

We ?rst replace Var [q(Y, X)] by its upper bound M 2 ? {E[q(Y, X)]}2 , ? which is tight when q takes n values in {?M, M }. Then inequality (2.5) is obtained by replacing n?1 i=1 q(Yi , Xi) in (2.6) by Zq and then bounding the suprema over the query space QA . In the de?nition of ?q in (2.5), we 4

incorporate the effect of uniform concentration over the query space QA by adding the quantity ?(QA), which measures the capacity of QA . In addition, under the De?nition 2.2, the algorithm A does not interact directly with data. Such an restriction characterizes the fact that in statistical problems, the effectiveness of an algorithm only depends on the global statistical properties, not the information of individual data points. For instance, algorithms that only rely on the convergence of the empirical distribution to the population distribution are contained in the statistical query model; whereas algorithms that hinge on the ?rst data point (y1 , x1) is not allowed. This restriction captures a vast family of algorithms in statistics and machine learning, including applying gradient method to maximize likelihood function, matrix factorization algorithms, expectation-maximization algorithms, and sampling algorithms [9]. Based on the statistical query model, we study the minimax risk under oracle complexity constraints. For the testing problem (2.3), let A(Tn) be a class of testing algorithms under the statistical query model with query complexity no more than Tn , with {Tn }n?1 being a sequence of positive integers depending on the sample size n. For any A ? A(Tn) and any oracle r ? R[?, n, Tn , ?(QA)] that responds to A , let H(A , r) be the set of test functions that deterministically depend on A ?s queries to the oracle r and the corresponding responses. We use P? to denote the distribution of the random variables returned by oracle r when the model parameter is ?. For a general hypothesis testing problem, namely, H0 : ? ? G0 versus H1 : ? ? G1 , the minimax testing risk with respect to an algorithm A and a statistical oracle r ? R[?, n, Tn , ?(QA)] is de?ned as ? ? ? Rn (G0 , G1 ; A , r) := inf sup P? (? = 1) + sup P? (? = 0) . (2.7) ??H(A ,r) ??G0

??G1

Compared with the classical minimax risk in (2.4), the new notion in (2.7) incorporates the computational budgets via oracle complexity. In speci?c, we only consider the test functions obtained by an algorithm with at most Tn queries

to a statistical oracle. If Tn is a polynomial of the dimensionality d, (2.7) characterizes the statistical optimality of computational ef?cient algorithms. This motivates us to de?ne the computationally tractable minimax rate, which contrasts with De?nition 2.1. De?nition 2.3 (Computationally tractable minimax rate). Let G1 (?n) := {? : ?(?) ? ?n } be a sequence of model spaces with minimum separation ?n , where ?(?) is the SNR. A sequence {? ?n }n?1 is called a computationally tractable minimax rate if ? For any sequence {?n }n?1 satisfying ?n = o(? ?n), any constant ? ¿ 0, and any A ? A(d?), ? there exists an oracle r ? R[?, n, Tn , ?(QA)] such that limn?? Rn [G0 , G1 (?n); A , r] = 1; ? ? For any sequence {?n }n?1 satisfying ?n = ?(? n), there exist a constant ? ¿ 0 and an algorithm ? A ? A(d?) such that, for any r ? R[?, n, Tn , ?(QA)], we have limn?? Rn [G0 , G1 (?n); A , r] = 0.

3

Main Results

Throughout this paper, we assume that the covariance matrix ? in (2.1) is known. Speci?cally, for some positive de?nite ? ? Rd?d , the parameter spaces of the null and alternative hypotheses are de?ned as G0 (?) := {? = (?, ?, ?, ?) : ? ? Rd },

(3.1) d

G1 (?; ?n) := {? = (?0 , ?1 , ?, ?) : ?0 , ?1 ? R , ?0 ? ?1 ? B0 (s), ?(?) ? ?n }.

Accordingly, the testing problem of detecting whether ?0 ?= ?1 is to distinguish H0 : ? ? G0 (?) versus H1 : ? ? G1 (?; ?n).

(3.2) (3.3)

In ?3.1, we present the minimax rate of the detection problem from an information-theoretic perspective. In ?3.2, under the statistical query model introduced in ?2.2, we provide a computational lower bound and a nearly matching upper bound that is achieved by an ef?cient testing algorithm. 3.1

Information-theoretic Limits

Now we turn to characterize the minimax rate given in De?nition 2.1. For?parameter spaces (3.1) and (3.2) with known ?, we show that in highly sparse setting where s = o(d), we have ? ?n? = s log d/n ? (1/?2 ? s log d/n), (3.4) 5

To prove (3.4), we ?rst present a lower bound which shows that the hypothesis testing problem in (3.3) is impossible if ?n = o(?n?). Theorem 3.1. For the hypothesis testing problem in (3.3) with known ?, we assume that there exists a small constant ? ¿ 0 such that s = o(d1/2??). Let ?n? be de?ned in (3.4). For any sequence {?n }n?1 such that ?n = o(?n?), any hypothesis test is asymptotically powerless, namely, lim sup Rn? [G0 (?), G1 (?; ?n)] = 1.

n??

?

By Theorem 3.1, we observe a phase transition in the necessary SNR for powerful detection when ? decreases from one to zero. Starting with rate s log d/n in the supervised setting where ? = 1, the required SNR gradually increases as label qualities decrease. Finally, when ? reaches? zero, which corresponds to the unsupervised setting, powerful detection requires the SNR to be ?(s log d/n). It is worth noting that when ? = (s log d/n)1/4 , we still have (n3 s

log d)1/4 uncorrupted labels. However, our lower bound (along with the upper bound shown in Theorem 3.2) indicates that the information contained in these uncorrupted labels are buried in the noise, and cannot essentially improve the detection quality compared with the unsupervised setting. Next we establish a matching upper bound for the detection problem in (3.3). We denote the condition number of the covariance matrix ? by ?, i.e., ? := ?max (?)/?min (?), where ?max (?) and ?min (?) are the largest and smallest eigenvalues of ?, repectively. Note that marginally Y is uniformly distributed over {0, 1}. For ease of presentation, we assume that the sample size is 2n and each class contains exactly n data points. Note that we can always discard some samples in the larger class to make the sample sizes of both classes to be equal. Due to the law of large numbers, this trick will not affect the analysis of sample complexity in the sense of order wise. d Given 2n i.i.d. samples {(yi , xi)}2n i=1 of (Y, X) ? {0, 1} ? R , we de?ne wi = ??1/2 (x2i ? x2i?1), for all i ? [n].

(0) n In addition, we split the dataset {(yi , xi)}2n i=1 into two disjoint parts {(0, xi)}i=1

and de?ne

(1)

ui = xi

(0)

? xi , for all i ? [n].

(3.5)

(1) and {(1, xi)}ni=1 ,

(3.6)

We note that computing sample differences in (3.5) and (3.6) is critical for our problem because we focus on detecting the difference between ?0 and ?1 , and computing differences can avoid estimating EP? (X) that might be dense. For any integer s ? [d], we de?ne B2 (s) := B0 (s) ? Sd?1 as the set of s-sparse vectors on the unit sphere in Rd . With {wi }ni=1 and {ui }ni=1 , we introduce two test functions ? ? n 1 ? (v? ??1 wi)2 ?1 := 1 sup ? 1 + ? , (3.7) 1 ? ?1 v v?B2 (s) n i=1 2v ? ? ? n 1? ?2 := 1 sup ?v, diag(?)?1/2 ui ? ? ?2 , (3.8) v?B2 (1) n i=1 where ?1 , ?2 ¿ 0 are algorithmic parameters that will be speci?ed later. To provide some intuitions, we consider the case where ? = I. Test function ?1 seeks a sparse direction that explains the most variance of wi . Therefore, such a test is closely related to the sparse principal component detection ?n problem [3]. Test function ?2 simply selects the coordinate of n?1 i=1 ui that has the largest magnitude and compares it with ?2 . This test is closely related to detecting sparse normal mean in high dimensions [16]. Based on these two ingredients, we construct our ?nal testing function ? as ? = ?1 ? ?2 , i.e., if any of ?1 and ?2 is true, then ? rejects the null. The following theorem establishes a suf?cient condition for test function ? to be asymptotically powerful. Theorem 3.2. Consider the testing problem (3.3) where ? is known and has condition number ?. For test functions ?1 and ?2 de?ned in (3.7) and (3.8) with parameters ?1 and ?2 given by ? ? ?1 = ? s log(ed/s)/n, ?2 = 8 log d/n.

We de?ne the ultimate test function as ? = ?1 ? ?2 . We assume that s ? C

? d for some absolute constant Cs and n ? 64 ? s log(ed/s). Then if ? ?n ? C ? ? ? [s log(ed/s)/n ? (1/?2 ? s log d/n)], (3.9) 6

where C ? is an absolute constant, then test function ? is asymptotically powerful. In speci?c, we have sup Pn? (? = 1) + ??G0 (?)

sup ??G1 (?;?n)

Pn? (? = 0) ? 20/d.

(3.10)

Theorem 3.2 provides a non-asymptotic guarantee. When n?goes to in?nity, (3.10) implies that the test function ? is asymptotically powerful. When s = o(d) and ? is a constant, (3.9) yields ? ?n = ?[s log d/n?(1/?2 ?s log d/n)], which matches the lower bound given in Theorem 3.1. Thus we conclude that ?n? de?ned in (3.4) is the minimax rate of testing problem in (3.3). We remark that when s = ?(d), ? = 1, i.e., the standard (low-dimensional) setting ? of two sample testing, the bound provided in (3.9) is sub-optimal as?[22] shows that SNR rate d/n is suf?cient for asymptotically powerful detection ? when n = ?(d). It is thus worth noting that we focus on the highly sparse setting s = o(d) and provided sharp minimax rate for ? ?this regime. In the de?nition of ?1 in (3.7), we search over the set B2 (s). Since B2 (s) contains ds distinct sets of supports, computing ?1 requires exponential running time. 3.2

Computational Limits

In this section, we characterize the computationally tractable minimax rate ? ?n given in De?nition 2.3. Moreover, we focus on the setting where ? is known a priori and the parameter spaces for the null and alternative hypotheses are de?ned ? in (3.1) and (3.2), respectively. The main result is that, in highly sparse setting where s = o(d), we have ? ? ?n = s2 /n ? (1/?2 ? s log d/n). (3.11)

We ?rst present the lower bound in the next result.

Theorem 3.3. For the testing problem in (3.3) with ? known a priori, we make the same assumptions as in Theorem 3.1. For any sequence {?n }n?1 such that ? ?? ?? s2 /n ? (1/?2 ? s/n) , (3.12) ?n = o ?n? ?

where ?n? is de?ned in (3.4), any computationally tractable test is asymptotically powerless under the statistical query model. That is, for any constant ? ¿ 0 and any A ? A(d?), there exists an oracle ? r ? R[?, n, Tn , ?(QA)] such that limn?? Rn [G0 (?), G1 (?, ?n); A , r] = 1.

We the lower bound in (3.12) differs from ?n? in (3.11) by a logarithmic term when ? remark 2that ? 1/n ? ? ? s log d/n. We expect this gap to be eliminated by more delicate analysis under the statistical query model. Now putting Theorems 3.1 and 3.3 together, we describe the ?more supervision, less computation? phenomenon as follows. (i) When 0 ? ? ? (log2 d/n)1/4 , the computational lower bound implies that the uncorrupted labels are unable to improve the quality of computationally tractable detection compared with the unsupervised setting. In addition, in this region, the gap between ?n? and ? ?n remains the same. (ii) When (log2 d/n)1/4 ¡ ? ? (s log d/n)1/4 , the information-theoretic lower bound shows that the uncorrupted labels cannot improve the quality of detection compared with unsupervised setting. However, more uncorrupted labels improve the statistical performances of hypothesis tests

that are computationally tractable by shrinking the gap between ?n? and ? ?n . (iii) When (s log d/n)1/4 ¡ ? ? 1, having more uncorrupted labels improves both statistical optimality and the computational ef?ciency. In speci?c, in this case, the gap between ?n? and ? ?n vanishes and we have ?n? = ? ?n = 1/?2 ? s log d/n. Now we derive a nearly matching upper bound under the statistical query model, which establishes the computationally tractable minimax rate together with Theorem 3.3. We construct a computationally ef?cient testing procedure that combines two test functions which yields the two parts in ? ?n respectively. Similar to ?1 de?ned in (3.7), the ?rst test function discards the information of labels, which works for the purely unsupervised setting where ? = 0. For j ? [d], we denote by ?j the j-th diagonal element of ?. Under the statistical query model, we consider the 2d query functions ? ? ? (3.13) qj (y, x) := xj / ?j ? 1{—xj / ?j — ? R ? log d}, ? ? q?j (y, x) := (x2j /?j ? 1) ? 1{—xj / ?j — ? R ? log d}, for all j ? [d], (3.14) 7

where R ¿ 0 is an absolute constant. Here we apply truncation to the query functions to obtain bounded queries, which is speci?ed by the statistical query model in De?nition 2.2. We denote by zqj and zq?j the realizations of the random variables output by the statistical oracle for query functions qj and q?j , respectively. As for the second test function, similar to (3.8), we consider ? ? ? q v (y, x) = (2y ? 1) ? v? diag(?)?1/2 x ? 1 —v? diag(?)?1/2 x— ? R ? log d (3.15)

for all v ? B2 (1). We denote by Zqv the output of the statistical oracle corresponding to query function q v . With these 4d query functions, we introduce test functions ? ? ? ? ?1 := 1 sup (zq?j ? zq2j) ? C? 1 , ?2 := 1 sup zqv ? 2? 2 , (3.16) j?[d]

v?B2 (1)

where ? 1 and ?2 are positive parameters that will be speci?ed later and C is an absolute constant. Theorem 3.4. For the test functions ?1 and ?2 de?ned in (3.16) , we de?ne the ultimate test function as ? = ?1 ? ?2 . We set ? ? ? ? 1 = R2 log d ? log(4d/?)/n, ? 2 = R log d ? log(4d/?)/n, (3.17) where ? = o(1). For the hypothesis testing problem in (3.3), we further assume that ??0 ?? ? ??1 ?? ? C0 for some constant C0 ¿ 0. Under the assumption that ? ?? ? ? sup (?0,j ? ?1,j)2 /?j = ? 1/?2 ? log2 d ? log(d/?)/n ? log d ? log(d/?)/n , (3.18) j?[d]

?

the risk of ? satis?es that Rn (?) = sup??G0 (?) P? (? = 1) + sup??G1 (?,?n) P? (? = 0) ? 5?. Here we denote by ?0,j and ?1,j the j-th entry of ?0 and ?1 , respectively. If we set the tail probability of the statistical query model to be ? = 1/d, (3.18) shows that ? is asymptotically powerful if supj?[d] (?0,j ? ?1,j)2 /?j = ?[(1/?2 ? log3 d/n) ? (log3 d/n)1/2]. ? When the energy of ?0 ? ?1 is spread over its support, ??0 ? ?1 ?? and ??0 ? ?1 ?2 / s are close. Under the assumption that the condition number ? of ? is a constant, (3.18) is implied by ?n ? (s2 log3 d/n)1/2 ? (1/?2 ? s log3 d/n).

Compared with Theorem 3.3, the above?upper bound matches the computational lower bound up to a logarithmic factor and ? ?n is between s2 /n ? (1/?2 ? s log d/n) and (s2 log3 d/n)1/2 ? (1/?2 ? s log3 d/n). Note that the

truncation on query functions in (3.13) and (3.14) yields an additional logarithmic term, which could be reduced to (s2 log d/n)1/2 ? (1/?2 ? s log d/n) using more delicate analysis. Moreover, the test function ?1 is essentially based on a diagonal thresholding algorithm performed on the covariance matrix ? of X. The work in [6] provides a more delicate analysis of this algorithm which establishes the s2 /n rate. Their algorithm can also be formulated into the statistical query model; we use the simpler version in (3.16) ? for ease of presentation. Therefore, with more sophicated proof techinique, it can be shown that s2 /n ? (1/?2 ? s log d/n) is the critical threshold for asymptotically powerful detection with computational ef?ciency. 3.3

Implication for Estimation

Our aforementioned phase transition in the detection problems directly implies the statistical and computational trade-offs in the problem of estimation. We consider the problem of estimating the parameter ?? = ?0 ? ?1 of the binary classi?cation model in (2.1) and (2.2), where ?? is s-sparse and ? is known a priori. We assume that the signal to noise ratio is ?(?) = ??? ??1 ?? ? ?n = o(? ?n). For any constant ? ¿ 0 and any A ? A(T) with T = O(d?), suppose we obtain an estimator ? of ?? by algorithm A under the statistical query model. If ?? ? converges to ?? in the sense that ?? ? ?1 ? ? (?? ? ??) ? (?? ? ??) = o[?n2 /?(?)],

? ? ??1 ?? ? ? ??? ??1 ??— = o(?n). Thus the test function ? = 1{?? ? ? ??? ? ? we have —?? ?n /2} is asymptotically powerful, which contradicts the computational lower bound in Theorem 3.3. ? ? ??)? ??1 (?? ? ? ??) ? C?n2 /?(?) for any Therefore, there exists a constant C such that (?? ? estimator ?? constructed from polynomial number of queries. Acknowledgments

We would like to thank Vitaly Feldman for valuable discussions. 8

2 References

[1] Angluin, D. and Laird, P. (1988). Learning from noisy examples. Machine Learning, 2 343?370. [2] Berthet, Q. and Rigollet, P. (2013). Computational lower bounds for sparse PCA. In Conference on Learning Theory. [3] Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. The Annals of Statistics, 41 1780?1815. [4] Chandrasekaran, V. and Jordan, M. I. (2013). Computational and statistical tradeoffs via convex relaxation. Proceedings of the National Academy of Sciences, 110 1181?1190. [5] Chen, Y. and Xu, J. (2014). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. arXiv preprint arXiv:1402.1267. [6] Deshpande, Y. and Montanari, A. (2014). Sparse PCA via covariance thresholding. In Advances in Neural Information Processing Systems. [7] Fan, J., Feng, Y. and Tong, X. (2012). A road to classi?cation in high dimensional space: The regularized optimal af?ne discriminant. Journal of the Royal Statistical Society: Series B, 74 745?771. [8] Fan, J., Liu, H., Wang, Z. and Yang, Z. (2016). Curse of heterogeneity: Computational barriers in sparse mixture models and phase re-

trieval. Manuscript. [9] Feldman, V., Grigorescu, E., Reyzin, L., Vempala, S. and Xiao, Y. (2013). Statistical algorithms and a lower bound for detecting planted cliques. In ACM Symposium on Theory of Computing. [10] Feldman, V., Guzman, C. and Vempala, S. (2015). Statistical query algorithms for stochastic convex optimization. arXiv preprint arXiv:1512.09170. [11] Feldman, V., Perkins, W. and Vempala, S. (2015). On the complexity of random satis?ability problems with planted solutions. In ACM Symposium on Theory of Computing. [12] Fr?nay, B. and Verleysen, M. (2014). Classi?cation in the presence of label noise: A survey. IEEE Transactions on Neural Networks and Learning Systems, 25 845?869. [13] Gao, C., Ma, Z. and Zhou, H. H. (2014). Sparse CCA: Adaptive estimation and computational barriers. arXiv preprint arXiv:1409.8565. [14] Garc?a-Garc?a, D. and Williamson, R. C. (2011). Degrees of supervision. In Advances in Neural Information Processing Systems. [15] Hajek, B., Wu, Y. and Xu, J. (2014). Computational lower bounds for community detection on random graphs. arXiv preprint arXiv:1406.6625. [16] Johnstone, I. M. (1994). On minimax estimation of a sparse normal mean vector. The Annals of Statistics, 22 271?289. [17] Joulin, A. and Bach, F. R. (2012). A convex relaxation for weakly supervised classi?ers. In International Conference on Machine Learning. [18] Kearns, M. (1993). Ef?cient noise-tolerant learning from statistical queries. In ACM Symposium on Theory of Computing. [19] Ma, Z. and Wu, Y. (2014). Computational barriers in minimax submatrix detection. The Annals of Statistics, 43 1089?1116. [20] Nettleton, D. F., Orriols-Puig, A. and Fornells, A. (2010). A study of the effect of different types of noise on the precision of supervised learning techniques. Arti?cial Intelligence Review, 33 275?306. [21] Patrini, G., Nielsen, F., Nock, R. and Carioni, M. (2016). Loss factorization, weakly supervised learning and label noise robustness. arXiv preprint arXiv:1602.02450. [22] Ramdas, A., Singh, A. and Wasserman, L. (2016). Classi?cation accuracy as a proxy for two sample testing. arXiv preprint arXiv:1602.02210. [23] Tony Cai, T., Liu, W. and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. Journal of the Royal Statistical Society: Series B, 76 349?372. [24] Tsybakov, A. B. (2008). Introduction to nonparametric estimation. Springer. [25] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. arXiv preprint arXiv:1011.3027. [26] Wang, T., Berthet, Q. and Samworth, R. J. (2014). Statistical and computational trade-offs in estimation of sparse principal components. arXiv preprint arXiv:1408.5369. [27] Wang, Z., Gu, Q. and Liu, H. (2015). Sharp computational-statistical phase transitions via oracle computational model. arXiv preprint arXiv:1512.08861. [28] Zhang, Y., Wainwright, M. J. and Jordan, M. I. (2014). Lower bounds on the performance of polynomialtime algorithms for sparse linear regression. In Conference on Learning Theory.

9