On Separability of Loss Functions, and Revisiting Discriminative Vs Generative Models

Authored by:

Pradeep K. Ravikumar
Adarsh Prasad
Alexandru Niculescu-Mizil

Abstract

We revisit the classical analysis of generative vs discriminative models for general exponential families, and high-dimensional settings. Towards this, we develop novel technical machinery, including a notion of separability of general loss functions, which allow us to provide a general framework to obtain l? convergence rates for general M-estimators. We use this machinery to analyze l? and l2 convergence rates of generative and discriminative models, and provide insights into their nuanced behaviors in high-dimensions. Our results are also applicable to differential parameter estimation, where the quantity of interest is the difference between generative model parameters.

1 Paper Body

Consider the classical conditional generative model setting, where we have a binary random response Y 2 {0, 1}, and a random covariate vector X 2 Rp , such that X—(Y = i) ? P?i for i 2 {0, 1}. Assuming that we know P (Y) and {P?i }1i=0 , we can use the Bayes rule to predict the response Y given covariates X. This is said to be the generative model approach to classification. Alternatively, consider the conditional distribution P (Y —X) as specified by the Bayes rule, also called the discriminative model corresponding to the generative model specified above. Learning this conditional model directly is said to be the discriminative model approach to classification. In a classical paper [8], the authors provided theoretical justification for the common wisdom regarding generative and discriminative models: when the generative model assumptions hold, the generative model estimators initially converge faster as a function of the number of samples, but have the same asymptotic error rate as discriminative models. And when the generative model assumptions do not hold, the discriminative model estimators eventually overtake the generative model estimators. Their analysis however was for the specific generative-discriminative

model pair of Naive Bayes, and logistic regression models, and moreover, was not under a high-dimensional sampling regime, when the number of samples could even be smaller than the number of parameters. In this paper, we aim to extend their analysis to these more general settings. Doing so however required some novel technical and conceptual developments. To motivate the machinery we develop, consider why the Naive Bayes model estimator might initially converge faster. The Naive Bayes model makes the conditional independence assumption that P (X—Y) = Qp s=1 P (Xs —Y), so that the parameters of each of the conditional distributions P (Xs —Y) for s 2 {1, . . . , p} could be estimated independently. The corresponding log-likelihood loss function is thus fully ?separable? into multiple components. The logistic regression log-likelihood on the other hand is seemingly much less ?separable?, and in particular, it does not split into multiple components each of which can be estimated independently. In general, we do not expect the loss functions underlying statistical estimators to be fully separable into multiple components, so that we need a more flexible notion of separability, where different losses could be shown to be separable to differing degrees. In a very related note, though it might seem unrelated at first, the analysis of ‘1 convergence rates of 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

statistical estimators considerably lags that of say ‘2 rates (see for instance, the unified framework of [7], which is suited to ‘2 rates but is highly sub-optimal for ‘1 rates). In part, the analysis of ‘1 rates is harder because it implicitly requires analysis at the level of individual coordinates of the parameter vector. While this is thus harder than an ‘2 error analysis, intuitively this would be much easier if the loss function were to split into independent components involving individual coordinates. While general loss functions might not be so ?fully separable?, they might perhaps satisfy a softer notion of separability motivated above. In a contribution that would be of independent interest, we develop precisely such a softer notion of separability for general loss functions. We then use this notion of separability to derive ‘1 convergence rates for general M -estimators.

Given this machinery, we are then able to contrast generative and discriminative models. We focus on the case where the generative models are specified by exponential family distributions, so that the corresponding discriminative models are logistic regression models with the generative model sufficient statistics as feature functions. To compare the convergence rates of the two models, we focus on the difference of the two generative model parameters, since this difference is also precisely the model parameter for the discriminative model counterpart of the generative model, via an application of the Bayes rule. Moreover, as Li et al. [3] and others show, the ‘2 convergence rates of the difference of the two parameters is what drives the classification error rates of both generative as well as discriminative model classifiers. Incidentally, such a difference of generative model parameters has also attracted interest outside the context of classification, where it is called differential parameter learning [1, 14, 6]. We thus analyze the ‘1 as well as ‘2 rates for both the generative and discriminative models, focusing on this parameter difference. As we show, unlike the case

of Naive Bayes and logistic regression in low-dimensions as studied in [8], this general highdimensional setting is more nuanced, and in particular depends on the separability of the generative models. As we show, under some conditions on the models, generative and discriminative models not only have potentially different ‘1 rates, but also differing ?burn in? periods in terms of the minimum number of samples required in order for the convergence rates to hold. The choice of a generative vs discriminative model, namely that with a better sample complexity, thus depends on their corresponding separabilities. As a minor note, we also show how generative model M -estimators are not directly suitable in high-dimensions, and provide a simple methodological fix in order to obtain better ‘2 rates. We instantiate our results with two running examples of isotropic and non-isotropic Gaussian generative models, and also corroborate our theory with instructive simulations.

2

Background and Setup.

We consider the problem of differential parameter estimation under the following generative model. Let Y 2 {0, 1} denote a binary response variable, and let X = (X1 , . . . , Xp) 2 Rp be the covariates. For simplicity, we assume P[Y = 1] = P[Y = 0] = 12 . We assume that conditioned on the response variable, the covariates belong to an exponential family, X—Y ? P?Y? (?), where: P?Y? (X—Y) = h(X) exp(h?Y? , (X)i

A(?Y?)).

(1)

Here, is the vector of the true canonical parameters, A(?) is the log-partition function and (X) (0) is the sufficient statistic. We assume access to two sets of samples X0n = {xi }ni=1 ? P?0? and ?Y?

(1)

X1n = {xi }ni=1 ? P?1? . Given these samples, as noted in the introduction, we are particularly ? interested in estimating the differential parameter ?diff := ?1? ?0? , since this is also the model parameter corresponding to the discriminative model, as we show below. In high dimensional ? ? sampling settings, we additionally assume that ?diff is at most s-sparse, i.e. ——?diff ——0 ? s.

We will be using the following two exponential family generative models as running examples: isotropic and non-isotropic multivariate Gaussian models. Isotropic Gaussians (IG) Let X = (X1 , . . . , Xp) ? N (?, Ip) be an isotropic gaussian random variable; it?s density can be written as: ? ? 1 1 T P? (x) = p exp (x ?) (x ?) . (2) 2 (2?)p

Gaussian MRF (GMRF). Let X = (X1 , . . . , Xp) denote a zero-mean gaussian random vector; 1 it?s density is fully-parametrized as by the inverse covariance or concentration matrix ? = (?) 0 2

and can be written as: P? (x) = r

1

? (2?)p det (?)

1

? exp

?
? 1 T x ?x . 2
(3)

Let d? = maxj2[p] ?(:,j) 0 is the maximum number non-zeros in a row of ?. Let ?? ? = Pp (??) 1 1 , where ———M ———1 is the ‘1 /‘1 operator norm given by ———M ———1 = max k=1 —Mjk —. j=1,2,...,p

Generative Model Estimation. Here, we proceed by estimating the two parameters {?i? }1i=0 individually. Letting ?b1 and ?b0 be the corresponding estimators, we can then estimate the difference of the parameters as ?bdiff = ?b1 ?b0 . The most popular class of estimators for the individual parameters is based on Maximum likelihood Estimation (MLE), where we maximize the likelihood of the given data. For isotropic gaussians, the negative log-likelihood function can be written as: ?T ? ?T ? b, (4) 2 P n where ? b = n1 i=1 xi . In the case of GGMs the negative log-likelihood function can be written as: DD EE b LnGGM (?) = ?, ? log(det(?)), (5) P P b = 1 n xi xT is the sample covariance matrix and hhU, V ii = where ? i i=1 i,j Uij Vij denotes n the trace inner product on the space of symmetric matrices. In high-dimensional sampling regimes (n ¡¡ p), regularized MLEs, for instance with ‘1 -regularization under the assumption of sparse model parameters, have been widely used [11, 10, 2]. Discriminative Model Estimation. Using Bayes rule, we have that: P[X—Y = 1]P[Y = 1] P[Y = 1—X] = P[X—Y = 0]P[Y = 0] + P[X—Y = 1]P[Y = 1] 1 = (6) ? 1 + exp ((h?1 ?0? , (x)i + c?)) where c? = A(?0?) A(?1?). The conditional distribution is simply a logistic regression model, with the generative model sufficient statistics as the features, and with optimal parameters being precisely ? the difference ?diff := ?1? ?0? of the generative model parameters. The corresponding negative log-likelihood function can be written as n 1X Llogistic (?, c) = (yi (h?, (xi)i + c) + (h?, (xi)i + c)) (7) n i=1 LnIG (?) =

where (t) = log(1 + exp(t)). In high dimensional sampling regimes, under the assumption that the model parameters are sparse, we would use the ‘1 -penalized version ?bdiff of the MLE (7) to estimate ? ?diff .

Outline. We proceed by studying the more general problem of ‘1 error for parameter estimation for any loss function Ln (?). Specifically, consider the general M -estimation problem, where we are given n i.i.d samples Z1n = {z1 , z2 , . . . , zn }, zi 2 Z from some distribution P, and we are interested in estimating some parameter ?? of the distribution P. Let ‘ : Rp ? Z 7! R be a twice differentiable and convex function which assigns a loss ‘(?; z) to any parameter ? 2 Rp , for a given ? observation z. Also assume that the loss is Fisher consistent so that ?? 2 argmin? L(?) where def ? L(?) = Ez?P [‘(?; z)] is the population loss. We are then interested in analyzing the M -estimators b ?? that minimize Pn the empirical loss i.e. ? 2 argmin? Ln (?), or regularized versions thereof, where Ln (?) = n1 i=1 L(?; Zi).

We introduce a notion of the separability of a loss function, and show how more separable losses require fewer samples to establish convergence for ?b ?? . We then instantiate our separability 1 results from this general setting for both generative and discriminative models. We calculate the number of samples

required for generative and discriminative approaches to estimate the differential ? parameter ?diff , for consistent convergence rates with respect to ‘1 and ‘2 norm. We also discuss the consequences of these results for high dimensional classification for Gaussian Generative models. 3

3

Separability

Let R(; ??) = rLn (?? +) rLn (??) r2 Ln (??) be the error in the first order approximation of the gradient at ?? . Let B1 (r) = {?— ——?——1 ? r} be an ‘1 ball of radius r. We begin by analyzing the low dimensional case, and then extend it to high dimensions. 3.1

Low Dimensional Sampling Regimes

In low dimensional sampling regimes, we assume that the number of samples n p. In this setting, we make the standard assumption that the empirical loss function Ln (?) is strongly convex. Let ?b = argmin? Ln (?) denote the unique minimizer of the empirical loss function. We begin by defining a notion of separability for any such empirical loss function Ln . Definition 1. Ln is (?, ,) locally separable around ?? if the remainder term R(; ??) satisfies: ——R(; ??)——1 ?

1

?

—— ——1 8

2 B1 ( )

This definition might seem a bit abstract, but for some general intuition, indicates the region where it is separable, ? indicates the conditioning of the loss, while it is that quantifies the degree of separability: the larger it is, the more separable the loss function. Next, we provide some additional intuition on how a loss function?s separability is connected to (?, ,). Using the mean-value theorem, we can write ——R(, ??)——1 = r2 Ln (?? + t) r2 Ln (??) for some t 2 (0, 1). This can 1 be further simplified as ——R(, ??)——1 ? r2 Ln (?? + t) r2 Ln (??) 1 —— ——1 . Hence, ? and 1/ measure the smoothness of Hessian (w.r.t. the ‘1 /‘1 matrix norm) in the neighborhood of ?? , with ? being the smoothness exponent, and 1/ being the smoothness constant. Note that the Hessian of the loss function r2 Ln (?) is a random matrix and can vary from being a diagonal matrix for a fully-separable loss function to a dense matrix for a heavily-coupled loss function. Moreover, from standard concentration arguments, the ‘1 /‘1 matrix norm for a diagonal ("separable") subgaussian random matrix has at most logarithmic dimension dependence1 , but for a dense ("non-separable") random matrix, the ‘1 /‘1 matrix norm could possibly scale linearly in the dimension. Thus, the scaling of ‘1 /‘1 matrix norm gives us an indication how ?separable? the matrix is. This intuition is captured by (?, ,), which we further elaborate in future sections by explicitly deriving (?, ,) for different loss functions and use them to derive ‘2 and ‘1 convergence rates. Theorem 1. Let Ln be a strongly convex loss function which is (?, ,) locally separable function ? 1 1 ? 1 ? 1} around ?? . Then, if ——rLn (??)——1 ? min{ 2? , 2?

where ? = r2 Ln (??)

1 1

.

?b

??

1

? 2? ——rLn (??)——1

Proof. (Proof Sketch). The proof begins by constructing a suitable continuous function F , for which b = ?b ?? is the unique fixed point. Next, we show that F (B1 (r)) ? B1 (r) for r = 2? ——rLn (??)——1 . Since F is continuous and ‘1 -ball is convex and compact, the contraction property coupled with Brouwer?s fixed point theorem [9], shows that there exists some fixed point of F , such that —— ——1 ? 2? ——rLn (??)——1 . By uniqueness of the fixed point, we then establish our result. See Figure 1 for a geometric description and Section A for more details 3.2

High Dimensional Sampling Regimes

In high dimensional sampling regimes (n ¡¡ p), estimation of model parameters is typically an under-determined problem. It is thus necessary to impose additional assumptions on the true model parameter ?? . We will focus on the popular assumption of sparsity, which entails that the number of non-zero coefficients of ?? is small, so that ——?? ——0 ? s. For this setting, we will be focusing in particular on ‘1 -regularized empirical loss minimization: 1

Follows from the concentration of subgaussian maxima [12]

4

F b

F(b) = b

F (B1 (2? ——rLn (??)——1))

B1 (2? ——rLn (??)——1)

Figure 1: Under the conditions of Theorem 1, F ( ) = r2 Ln (??) 1 (R(; ??) + rLn (??)) is contractive over B1 (2? ——rLn (??)——1) and has b = ?b ?? as its unique fixed point. Using these two observations, we can conclude that b ? 2? ——rLn (??)—— . 1

1

?b

n

= argmin Ln (?) + ?

n

(8)

——?——1

Let S = {i — ?i? 6= 0} be the support set of the true parameter and M(S) = {v—vS c = 0} be the corresponding subspace. Note that under a high-dimensional sampling regime, we can no longer assume that the empirical loss Ln (?) is strongly convex. Accordingly, we make the following set of assumptions: ? Assumption 1 (A1): Positive Definite Restricted Hessian. r2SS Ln (??) % min I ? Assumption 2 (A2): Irrepresentability. There exists some 2 (0, 1] such that r2S c S Ln (??) r2SS Ln (??)

1

1
?1
? Assumption 3 (A3). Unique Minimizer. When restricted to the true support, the solution to the ‘1 penalized loss minimization problem is unique, which we denote by: ?? = argmin {Ln (?) + n ——?—— } . (9) n
1
?2M(S)
Assumptions 1 and 2 are common in high dimensional analysis. We verify that Assumption 3 holds for different loss functions individually. We refer the reader to [13, 5, 11, 10] for further details on these assumptions. For this high dimensional sampling regime, we also modify our separability notion to a restricted separability, which entails that the remainder term be separable only over the model subspace M(S). Definition 2. Ln is (?, ,) restricted locally separable around ?? over the subspace M(S) if the remainder term R(; ??) satisfies: ——R(; ??)——1 ?
1
?
—— ——1 8
2 B1 ( ) M(S)
We present our main deterministic result in high dimensions. Theorem 2. Let Ln be a (?, ,) locally separable function around ?? . If (that, ?
8
n
——rLn (??)——1 .
? ——rLn (??)——1 +
n
? min
n
2? ,
1 2?
? ? 1
1 ?
1
o
Then we have that support(?b n) ? support(??) and where ? = r2SS Ln (??)
1
?b
n
??
1
? 2? (——rLn (??)——1 +
1
5
n)
n , rLn (?

?

)) are such

Proof. (Proof Sketch). The proof invokes the primal-dual witness argument [13] which when combined with Assumption 1-3, gives ?b n 2 M(S) and that ?b n is the unique solution of the restricted problem. The rest of the proof proceeds similar to Theorem 1, by constructing a suitable function F : R—S— 7! R—S— for which b = ?b n ?? is the unique fixed point, and showing that F is contractive over B1 (r; ??) for r = 2? (——rLn (??)——1 + n).See Section B for more details. Discussion. Theorems 1 and 2 provide a general recipe to estimate the number of samples required by any loss ‘(?, z) to establish ‘1 convergence. The first step is to calculate the separability constants (?, ,) for the corresponding empirical loss function Ln . Next, since the loss ‘ is Fisher consistent, ? ?) = 0, the upper bound on ——rLn (??)—— can be shown to hold by analyzing the so that rL(? 1 concentration of rLn (??) around its mean. We emphasize that we do not impose any restrictions on the values of (?, ,). In particular, these can scale with the number of samples n; our results hold so long as the number of samples n satisfy the conditions of the theorem. As a rule of thumb, the smaller that either or get for any given loss ‘, the larger the required number of samples.

4

‘1 -rates for Generative and Discriminative Model Estimation

In this section we study the ‘1 rates for differential parameter estimation for the discriminative and generative approaches. We do so by calculating the separability of discriminative and generative loss functions, and then instantiate our previously derived results. 4.1

Discriminative Estimation

As discussed before, the discriminative approach uses ‘1 -regularized logistic regression with the sufficient statistic as features to estimate the differential parameter. In addition to A1-A3, we Pn 2 assume column normalization of the sufficient statistics, i.e. i=1 ([(xi)]j) ? n. Let n = maxi —— (x)i ——1 , ?n = maxi ——((x)i)S ——2 . Firstly, we characterize the separability of the logistic loss. ? ? Lemma 1. The logistic regression negative log-likelihood LnLogistic from (7) is 2, s n1? 2 , 1 ren stricted local separable around ?? .

Combining Lemma 1 with Theorem 2, we get the following corollary. Corollary 3. (Logistic Regression) Consider the model in (1), then there exist q universal positive constants C1 , C2 and C3 such that for n differential estimate ?bdiff , satisfies 4.2

support(?bdiff) ?

C 1 ?2 s 2

? support(?diff)

2 4 n ?n

and

Generative Estimation

log p and

?bdiff

? ?diff

n
= C2
1
? C3
log p n ,
r
the discriminative
log p . n

We characterize the separability of Generative Exponential Families. The negative log-likelihood function can be written as: Ln (?) = A(?) h?, n i , P n where n = n1 i=1 (xi). In this setting, the remainder term is independent of the data and can be written as R( ) = rA(?? +) rA(??) r2 A(??) and rLn (??) = E[(x)] n1 (xi). Hence, ——rLn (??)——1 is a measure of how well the sufficient statistics concentrate around their mean. Next, we show the separability of our running examples Isotropic Gaussians and Gaussian Graphical Models. Lemma 2. The isotropic Gaussian negative log-likelihood LnIG from (4) is (?, 1, 1) locally separable around ?? . ? ? Lemma 3. The Gaussian MRF negative log-likelihood LnGGM from (5) is 2, 3d? 2?3 , 3d? 1??? ? ? ?? restricted locally separable around ?? . 6

Comparing Lemmas 1, 2 and 3, we see that the separability of the discriminative model loss depends weakly on the feature functions. On the other hand, the separability for the generative model loss depends critically on the underlying sufficient statistics. This has consequences for their differing sample complexities for differential parameter estimation, as we show next. Corollary 4. (Isotropic Gaussians) Consider the model in (2). Then there exist universal constants C1 , C2 , C3 such that if the number of samples scale as n C1 log p, then with probability atleast 1 1/pC2 , the generative estimate of the differential parameter ?bdiff satisfies r log p ? b ?diff ?diff ? C3 . n 1

Comparing Corollary 3 and Corollary 4, we see that for isotropic gaussians, both the discriminative and generative approach achieve the same ‘1 convergence rates, but at different sample complexities. Specifically, the sample complexity for the generative method depends only logarithmically on the dimension p, and is independent of the differential sparsity s, while the sample complexity of the discriminative method depends on the differential sparsity s. Therefore in this case, the generative method is strictly better than its discriminative counterpart, assuming that the generative model assumptions hold. Corollaryp5. (Gaussian MRF) Consider the model in (3), and suppose that the scaled covari? are subgaussian with parameter 2 . Then there exist universal positive conates Xk / ?kk stants C2 , C3 , C4 such that if the number of samples for the two generative models scale as ni C2 ?2i ?6(??) 1 d2?? log p, for i 2 {0, 1}, then with probability at least 1 1/pC3 , the geni i bdiff = ? b1 ? b0 , satisfies erative estimate of the differential parameter, ? r log p ? bdiff ?diff ? ? C4 , n 1 bi) ? support(??) for i 2 {0, 1}. and support(? i

Comparing Corollary 3 and Corollary 5, we see that for Gaussian Graphical Models, both the discriminative and generative approach achieve the same ‘1 convergence rates, but at different sample complexities. Specifically, the sample

complexity for the generative method depends only on row-wise sparsity of the individual models d2?? , and is independent of sparsity s of the differential i ? parameter ?diff . In contrast, the sample complexity of the discriminative method depends only on the sparsity of the differential parameter, and is independent of the structural complexities of the individual model parameters. This suggests that in high dimensions, even when the generative model assumptions hold, generative methods might perform poorly if the underlying model is highly non-separable (e.g. d = ?(p)), which is in contrast to the conventional wisdom in low dimensions.

Related Work. Note that results similar to Corollaries 3 and 5 have been previously reported in [11, 5] separately. Under the same set of assumptions as ours, Li et al. [5] provide a unified analysis for support recovery and ‘1 -bounds for ‘1 -regularized M-estimators. While they obtain the same rates as ours, their required sample complexities are much higher, since they do not exploit the separability of the underlying loss function. As one example, in the case of GMRFs, their results require the number of samples to scale as n ¿ k 2 log p, where k is the total number of edges in the graph, which is sub-optimal, and in particular does not match the GMRF-specific analysis of [11]. On the other hand, our unified analysis is tighter, and in particular, does match the results of [11].

5

‘2 -rates for Generative and Discriminative Model Estimation

In this section we study the ‘2 rates for differential parameter estimation for the discriminative and generative approaches. 5.1

Discriminative Approach

The bounds for the discriminative approach are relatively straightforward. Corollary 3 gives bounds b ? support(??). Since the true model parameter is on the ‘1 error and establishes that support(?) p s-sparse, ——?? ——0 ? s, the ‘2 error can be simply bounded as s k?b ?? k1 . 7

5.2

Generative Approach

In the previous section, we saw that the generative approach is able to exploit the inherent separability of the underlying model, and thus is able to get ‘1 rates for differential parameter estimation at a much lower sample complexity. Unfortunately, it does not q have support consistency. Hence a na?ve p generative estimator will have an ‘2 error scaling with p log n , which in high dimensions, would ? make it unappealing. However, one can exploit the sparsity of ?diff and get better rates of convergence in ‘2 -norm by simply soft-thresholding the generative estimate. Moreover, soft-thresholding also leads to support consistency. Definition 3. We denote the soft-thresholding operator ST ST

n

(?) = argmin w

1 ——w 2

n

(?), defined as:

2

?——2 +
n
——w——1 .

Lemma 4. Suppose ? = ?? + ? for some s-sparse ?? . Then there exists a universal constant C1 such that for n 2 ——?——1 , p ——ST n (?) ?? ——2 ? C1 s ——?——1 and ——ST n (?) ?? ——1 ? C1 s ——?——1 (10) Note that this is a completely deterministic result and has no sample complexity requirement. Motivated by this, we introduce a thresholded generative estimator that has two stages: (a) compute ?bdiff using the generative model estimates, and (b) soft-threshold the generative estimate with n = ? c ?bdiff ?diff . An elementary application of Lemma 4 can then be shown to provide ‘2 error 1

? bounds for ?bdiff given its ‘1 error bounds, and that the true parameter ?diff is s-sparse. We instantiate these ‘2 -bounds via corollaries for our running examples of Isotropic Gaussians, and Gaussian MRFs.

Lemma 5. (Isotropic Gaussians) Consider the model in (2). Then there exist universal constants C1 , C2 , C3 such that if the number of samples scale as n C1 log p, then with probability ? ?atleast 1 1/pC2 , the soft-thresholded generative estimate of the differential parameter ST n ?bdiff , with q the soft-thresholding parameter set as n = c logn p for some constant c, satisfies: r ? ? s log p ? b ST n ?diff ?diff ? C3 . n 2

Lemma 6. MRF) Consider the model in Equation 3, and suppose that the covarip (Gaussian ? are subgaussian with parameter 2 . Then there exist universal positive conates Xk / ?kk stants C2 , C3 , C4 such that if the number of samples for the two generative models scale as ni C2 ?2i ?6(??) 1 d2?? log p, for i 2 {0, 1}, for i 2 {0, 1}, then with probability at least 1 1/pC3 , i i ? ? bdiff , with the softthe soft-thresholded generative estimate of the differential parameter, ST n ? q thresholding parameter set as n = c logn p for some constant c, satisfies: r ? ? s log p ? bdiff ST n ? ?diff ? C4 . n 2 Comparing Lemmas 5 and 6 to Section 5.1, we can see that the additional soft-thresholding step allows the generative methods to achieve the same ‘2 -error rates as the discriminative methods, but at different sample complexities. The sample complexities of the generative estimates depend on the separabilities of the individual models, and is independent of the differential sparsity s, where as the sample complexity of the discriminative estimate depends only on the differential sparsity s.

6

Experiments: High Dimensional Classification

In this section, we corroborate our theoretical results on ‘2 -error rates for generative and discriminative model estimators, via their consequences for high dimensional classification. We focus on the case of isotropic Gaussian generative models X—Y ? N (?Y , Ip), where ?0 , ?1 2 Rp are unknown 8

0-1 Error for s=4,p=512,d=1

0.5

0.4

0-1 Error

0.3 0.25

0.35 0.3 0.25
0.2
50
100
150
200
250
300
350
400
0.15
0.3
0.2
0
50
100
150
200
n
(a)
0.35
0.25
0.2
0
Gen-Thresh Logistic
0.45
0.4
0-1 Error
0-1 Error
0.4
0-1 Error for s=64,p=512,d=1
0.5 Gen-Thresh Logistic
0.45
0.35
0.15
0-1 Error for s=16,p=512,d=1
0.5 Gen-Thresh Logistic
0.45
250
300
350
400
0.15
0
50
100

n
(b)
s = 4, p = 512
150
200
250
300
350
400
n
(c)
s = 16, p = 512
Figure 2: Effect of sparsity s on excess 0
s = 64, p = 512
1 error.

and ?1 ?0 is s-sparse. Here, we are interested in a classifier C : Rp 7! {0, 1} that achieves low classification error EX,Y [1 {C(X) 6= Y }]. Under this setting, it can be shown that the Bayes classifier, that achieves the lowest possible classification error, is given by the linear discriminant ?T ? ?T ? classifier C ? (x) = 1 xT w? + b? ¿ 0 , where w? = (?1 ?0) and b? = 0 0 2 1 1 . Thus, the coefficient w? of the linear discriminant is precisely the differential parameter, which can be estimated via both generative and discriminative approaches as detailed in the previous section. Moreover, the classification error can also be related to the ‘2 error of thenestimates. Under o some mild assumptions, T b b Li et al. [3] showed that for any linear classifier C(x) = 1 x w b + b ¿ 0 , the excess classification error can be bounded as: ? b ? C1 ——w E(C) b

2 w? ——2 + bb
b?
2 2
?
,

for some constant C1 ¿ 0, and where E(C) = EX,Y [1 {C(X) 6= Y }] EX,Y [1 {C ? (X) 6= Y }] is the excess 0-1 error. In other words, the excess classification error is bounded by a constant times the ‘2 error of the differential parameter estimate. Methods. In this setting, as discussed in previous sections, the discriminative model is simply a logistic regression model with linear features (6), so that the discriminative estimate of the differential parameter w b as well as the constant bias term bb can be simply obtained via ‘1 -regularized logistic regression. For the generative estimate, we use our two stage estimator from Section 5, which proceeds by estimating ? b0 , ? b1 using the empirical means, and then estimating the differential parameter by soft-thresholding q the difference of the generative model parameter estimates w bT = ST n (b ?1 ? b0)

where ?bT =
n
1 2

= C1
log p n
for some constant C1 . The corresponding estimate for b? is given by
hw bT , ? b1 + ? b0 i.

Experimental Setup.? For our experimental? setup, we consider isotropic Gaussian models with 1s 1s means ?0 = 1p p1s , ?1 = 1p + p1s , and vary the sparsity level s. For both methods, 0p s 0p s p we set the regularization parameter 2 as n = log(p)/n. We report the excess classification error for the two approaches, averaged over 20 trials, in Figure 2. Results. As can be seen from Figure 2, our two-staged thresholded generative estimator is always better than the discriminative estimator, across different sparsity levels s. Moreover, the sample complexity or ?burn-in? period of the discriminative classifier strongly depends on the sparsity level, which makes it unsuitable when the true parameter is not highly sparse. For our two-staged generative estimator, we see that the sparsity s has no effect on the ?burn-in? period of the classifier. These observations validate our theoretical results from Section 5. 2

See Appendix J for cross-validated plots.

9

Acknowledgements A.P. and P.R. acknowledge the support of ARO via W911NF-12-1-0390 and NSF via IIS-1149803, IIS-1447574, DMS-1264033, and NIH via R01 GM117594-01 as part of the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences.

2 References

[1] Alberto de la Fuente. From ?differential expression?to ?differential networking??identification of dysfunctional regulatory networks in diseases. Trends in genetics, 26(7):326?333, 2010. [2] Christophe Giraud. Introduction to high-dimensional statistics, volume 138. CRC Press, 2014. [3] Tianyang Li, Adarsh Prasad, and Pradeep K Ravikumar. Fast classification rates for high-dimensional gaussian generative models. In Advances in Neural Information Processing Systems, pages 1054?1062, 2015. [4] Tianyang Li, Xinyang Yi, Constantine Carmanis, and Pradeep Ravikumar. Minimax gaussian classification & clustering. In Artificial Intelligence and Statistics, pages 1?9, 2017. [5] Yen-Huan Li, Jonathan Scarlett, Pradeep Ravikumar, and Volkan Cevher. Sparsistency of 1-regularized m-estimators. In AISTATS, 2015. [6] Song Liu, John A Quinn, Michael U Gutmann, Taiji Suzuki, and Masashi Sugiyama. Direct learning of sparse changes in markov networks by density ratio estimation. Neural computation, 26(6):1169?1197, 2014. [7] Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for highdimensional analysis of m-estimators with decomposable regularizers. In Advances in Neural Information Processing Systems, pages 1348?1356, 2009. [8] Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. Advances in neural information processing systems, 2:841?848, 2002. [9] James M Ortega and Werner

C Rheinboldt. Iterative solution of nonlinear equations in several variables. SIAM, 2000. [10] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ‘1-regularized logistic regression. The Annals of Statistics, 38(3):1287?1319, 2010. [11] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ‘1 -penalized log-determinant divergence. Electronic Journal of Statistics, 5: 935?980, 2011. [12] JM Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. preparation. University of California, Berkeley, 2015. [13] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (lasso). IEEE transactions on information theory, 55(5):2183?2202, 2009. [14] Sihai Dave Zhao, T Tony Cai, and Hongzhe Li. Direct estimation of differential networks. Biometrika, page asu009, 2014.

10