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

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 $\ell_\infty$ convergence rates for general $M$-estimators. We use this machinery to analyze $\ell_\infty$ and $\ell_2$ 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 Introduction

Consider the classical conditional generative model setting, where we have a binary random response $Y \in \{0,1\}$, and a random covariate vector $X \in \mathbb{R}^p$, such that $X|Y = i \sim P_{\theta_i}$ for $i \in \{0,1\}$. Assuming that we know $P(Y)$ and $\{P_{\theta_i}\}_{i=0}^1$, 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 $P(X|Y) = \prod_{s=1}^p P(X_s|Y)$, so that the parameters of each of the conditional distributions $P(X_s|Y)$ for $s \in \{1,\ldots,p\}$ 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 $\ell_\infty$ convergence rates of
We will be using the following two exponential family generative models as running examples:

We consider the problem of differential parameter estimation under the following generative model. While this is thus harder than an $\ell_2$ 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 $\ell_\infty$ 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 variable, the covariates belong to an exponential family, via an application of the Bayes rule. Moreover, as Li et al. [3] and others show, the vector. While this is thus harder than an $\ell_2$ convergence 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 $Y \in \{0, 1\}$ denote a binary response variable, and let $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ be the covariates. For simplicity, we assume $P[Y = 1] = P[Y = 0] = \frac{1}{2}$. We assume that conditioned on the response variable, the covariates belong to an exponential family, $X|Y \sim P_{\theta^*_Y}(\cdot)$, where:

$$P_{\theta^*_Y}(X|Y) = h(X) \exp\langle \theta^*_Y, \phi(X) \rangle - A(\theta^*_Y). \quad (1)$$

Here, $\theta^*_Y$ is the vector of the true canonical parameters, $A(\theta)$ is the log-partition function and $\phi(X)$ is the sufficient statistic. We assume access to two sets of samples $X^{(1)}_n = \{x^{(0)}_i\}_{i=1}^n \sim P_{\theta^*_0}$ and $X^{(1)}_n = \{x^{(1)}_i\}_{i=1}^n \sim P_{\theta^*_1}$. Given these samples, as noted in the introduction, we are particularly interested in estimating the differential parameter $\theta^*_\text{diff} := \theta^*_1 - \theta^*_0$, 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 $\theta^*_\text{diff}$ is at most $s$-sparse, i.e. $\|\theta^*_\text{diff}\|_0 \leq s$.

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 $X = (X_1, \ldots, X_p) \sim \mathcal{N}(\mu, \Sigma_p)$ be an an isotropic gaussian random variable; its density can be written as:

$$P_{\mu}(x) = \frac{1}{\sqrt{(2\pi)^p}} \exp \left( -\frac{1}{2}(x - \mu)^T(x - \mu) \right). \quad (2)$$

**Gaussian MRF (GMRF).** Let $X = (X_1, \ldots, X_p)$ denote a zero-mean gaussian random vector; its density is fully-parametrized as by the inverse covariance or concentration matrix $\Theta = (\Sigma)^{-1} \succ 0$
We introduce a notion of the separability of a loss function, and show how more separable losses can be written as:

\[ \mathbb{P}_\theta(x) = \frac{1}{\sqrt{2\pi}^d}\exp\left(-\frac{1}{2}x^T\Theta x\right). \]

Let \( d_\theta = \max_{j \in [p]} \|\Theta_{i,j}\|_1 \) be the maximum number non-zeros in a row of \( \Theta \). Let \( \kappa_\ell^* = \| (\Theta^*)^{-1} \|_\infty \), where \( \| M \|_\infty \) is the \( \ell_\infty/\ell_\infty \) operator norm given by \( \| M \|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^p |M_{jk}| \).

**Generative Model Estimation.** Here, we proceed by estimating the two parameters \( \{\theta^*_1\}_{i=0}^1 \) individually. Letting \( \hat{\theta}_1 \) and \( \hat{\theta}_0 \) be the corresponding estimators, we can then estimate the difference of the parameters as \( \hat{\theta}_{\text{diff}} = \hat{\theta}_1 - \hat{\theta}_0 \). 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:

\[ \mathcal{L}_{\text{nll}}(\theta) = \frac{\theta^T \theta}{2} - \theta^T \bar{\mu}, \]

where \( \bar{\mu} = \frac{1}{n} \sum_{i=1}^n x_i \). In the case of GGMs the negative log-likelihood function can be written as:

\[ \mathcal{L}_{\text{nll}}(\theta) = \frac{1}{n} \sum_{i=1}^n \log(|\Sigma_i|) - \frac{1}{2} \text{tr}(\Sigma_i^{-1} \bar{y}_i) + \frac{1}{2} \text{tr}(\Sigma_i^{-1} \bar{y}_i \Sigma_i^{-1}) \]

where \( \bar{y}_i = \frac{1}{p} \sum_{j=1}^p x_i \), and \( \Sigma_i \) is the sample covariance matrix and \( \|\Theta\|_2 \) denotes the trace inner product on the space of symmetric matrices. In high-dimensional sampling regimes \( (n << p) \), regularized MLEs, for instance with \( \ell_1 \)-regularization under the assumption of sparse model parameters, have been widely used [11, 10, 2].

**Discriminative Model Estimation.** Using Bayes rule, we have that:

\[ \mathbb{P}[Y = 1 | X] = \frac{\mathbb{P}[X | Y = 1] \mathbb{P}[Y = 1]}{\mathbb{P}[X | Y = 0] \mathbb{P}[Y = 0] + \mathbb{P}[X | Y = 1] \mathbb{P}[Y = 1]} = \frac{1}{1 + \exp(-((\theta^*_1 - \theta^*_0, \phi(x)) + c^*))} \]

where \( c^* = A(\theta^*_1) - A(\theta^*_0) \). 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 \( \theta_{\text{diff}}^* := \theta^*_1 - \theta^*_0 \) of the generative model parameters. The corresponding negative log-likelihood function can be written as:

\[ \mathcal{L}_{\text{logistic}}(\theta, c) = \frac{1}{n} \sum_{i=1}^n (-y_i((\theta, \phi(x_i)) + c) + \Phi((\theta, \phi(x_i)) + c)) \]

where \( \Phi(t) = \log(1 + \exp(t)) \). In high dimensional sampling regimes, under the assumption that the model parameters are sparse, we would use the \( \ell_1 \)-penalized version \( \hat{\theta}_{\text{diff}} \) of the MLE (7) to estimate \( \theta_{\text{diff}}^* \).

**Outline.** We proceed by studying the more general problem of \( \ell_\infty \) error for parameter estimation for any loss function \( \mathcal{L}_n(\cdot) \). Specifically, consider the general \( M \)-estimation problem, where we are given \( n \) i.i.d samples \( Z^n = \{z_1, z_2, \ldots, z_n\}, z_i \in \mathcal{Z} \) from some distribution \( \mathbb{P} \), and we are interested in estimating some parameter \( \theta^* \) of the distribution \( \mathbb{P} \). Let \( \ell : \mathbb{R}^p \times \mathcal{Z} \rightarrow \mathbb{R} \) be a twice differentiable and convex function which assigns a loss \( \ell(\theta; z) \) to any parameter \( \theta \in \mathbb{R}^p \), for a given observation \( z \). Also assume that the loss is Fisher consistent so that \( \theta^* \in \arg\min_{\theta \in \mathbb{R}^p} \mathcal{L}(\theta) \) where \( \mathcal{L}(\theta) \equiv \mathbb{E}_{z \sim \mathbb{P}}[\ell(\theta; z)] \) is the population loss. We are then interested in analyzing the \( M \)-estimators \( \theta^* \) that minimize the empirical loss i.e. \( \hat{\theta} \in \arg\min_{\theta} \mathcal{L}_n(\theta) \), or regularized versions thereof, where \( \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\theta; Z_i) \).

We introduce a notion of the separability of a loss function, and show how more separable losses require fewer samples to establish convergence for \( \|\hat{\theta} - \theta^*\|_\infty \). We then instantiate our separability 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 \( \theta_{\text{diff}}^* \), for consistent convergence rates with respect to \( \ell_\infty \) and \( \ell_2 \) norm. We also discuss the consequences of these results for high dimensional classification for Gaussian Generative models.
3 Separability

Let $R(\Delta; \theta^*) = \nabla \mathcal{L}_n(\theta^* + \Delta) - \nabla \mathcal{L}_n(\theta^*) - \nabla^2 \mathcal{L}_n(\theta^*) \Delta$ be the error in the first order approximation of the gradient at $\theta^*$. Let $B_{\infty}(r) = \{ \theta | \| \theta \|_{\infty} \leq r \}$ be an $\ell_\infty$ 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 $n \gg p$. In this setting, we make the standard assumption that the empirical loss function $\mathcal{L}_n(\cdot)$ is strongly convex.

Let $\hat{\theta} = \text{argmin}_\theta \mathcal{L}_n(\theta)$ denote the unique minimizer of the empirical loss function. We begin by defining a notion of separability for any such empirical loss function $\mathcal{L}_n$.

**Definition 1.** $\mathcal{L}_n$ is $(\alpha, \beta, \gamma)$ locally separable around $\theta^*$ if the remainder term $R(\Delta; \theta^*)$ satisfies:

$$\| R(\Delta; \theta^*) \|_{\infty} \leq \frac{1}{\beta} \| \Delta \|_{\infty} \forall \Delta \in B_{\infty}(\gamma)$$

This definition might seem a bit abstract, but for some general intuition, $\gamma$ indicates the region where it is separable, $\alpha$ indicates the conditioning of the loss, while it is $\beta$ 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 $(\alpha, \beta, \gamma)$. Using the mean-value theorem, we can write $\| R(\Delta, \theta^*) \|_{\infty} = \| (\nabla^2 \mathcal{L}_n(\theta^* + t\Delta) - \nabla^2 \mathcal{L}_n(\theta^*)) \Delta \|_{\infty}$ for some $t \in (0, 1)$. This can be further simplified as $\| R(\Delta, \theta^*) \|_{\infty} \leq \| \nabla^2 \mathcal{L}_n(\theta^* + t\Delta) - \nabla^2 \mathcal{L}_n(\theta^*) \|_{\infty} \| \Delta \|_{\infty}$. Hence, $\alpha$, $\beta$, and $1/\beta$ measure the smoothness of Hessian (w.r.t. the $\ell_\infty/\ell_\infty$ matrix norm) in the neighborhood of $\theta^*$, with $\alpha$ being the smoothness exponent, and $1/\beta$ being the smoothness constant. Note that the Hessian of the loss function $\nabla^2 \mathcal{L}_n(\theta)$ 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 $\ell_\infty/\ell_\infty$ matrix norm for a diagonal (“separable”) subgaussian random matrix has at most logarithmic dimension dependence\footnote{Follows from the concentration of subgaussian maxima [12]}, but for a dense (“non-separable”) random matrix, the $\ell_\infty/\ell_\infty$ matrix norm could possibly scale linearly in the dimension. Thus, the scaling of $\ell_\infty/\ell_\infty$ matrix norm gives us an indication how “separable” the matrix is. This intuition is captured by $(\alpha, \beta, \gamma)$, which we further elaborate in future sections by explicitly deriving $(\alpha, \beta, \gamma)$ for different loss functions and use them to derive $\ell_2$ and $\ell_\infty$ convergence rates.

**Theorem 1.** Let $\mathcal{L}_n$ be a strongly convex loss function which is $(\alpha, \beta, \gamma)$ locally separable function around $\theta^*$. Then, if $\| \nabla \mathcal{L}_n(\theta^*) \|_{\infty} \leq \min\{ \frac{\gamma}{\beta}, \frac{1}{\beta^{1-\alpha}} \}$

$$\| \hat{\theta} - \theta^* \|_{\infty} \leq 2\kappa \| \nabla \mathcal{L}_n(\theta^*) \|_{\infty}$$

where $\kappa = \| \nabla^2 \mathcal{L}_n(\theta^*)^{-1} \|_{\infty}$.

**Proof.** (Proof Sketch). The proof begins by constructing a suitable continuous function $F$, for which $\hat{\Delta} = \hat{\theta} - \theta^*$ is the unique fixed point. Next, we show that $F(B_{\infty}(r)) \subseteq B_{\infty}(r)$ for $r = 2\kappa \| \nabla \mathcal{L}_n(\theta^*) \|_{\infty}$. Since $F$ is continuous and $\ell_\infty$-ball is convex and compact, the contraction property coupled with Brouwer’s fixed point theorem [9], shows that there exists some fixed point $\Delta$ of $F$, such that $|\Delta|_{\infty} \leq 2\kappa \| \nabla \mathcal{L}_n(\theta^*) \|_{\infty}$. 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 ($n \ll p$), estimation of model parameters is typically an under-determined problem. It is thus necessary to impose additional assumptions on the true model parameter $\theta^*$. We will focus on the popular assumption of sparsity, which entails that the number of non-zero coefficients of $\theta^*$ is small, so that $|\theta^*|_0 \leq s$. For this setting, we will be focusing in particular on $\ell_1$-regularized empirical loss minimization.
Then we have that

We present our main deterministic result in high dimensions. 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.

Let $S = \{i | \theta^*_i \neq 0\}$ be the support set of the true parameter and $\mathcal{M}(S) = \{v | v_S = 0\}$ be the corresponding subspace. Note that under a high-dimensional sampling regime, we can no longer assume that the empirical loss $\mathcal{L}_n(\cdot)$ is strongly convex. Accordingly, we make the following set of assumptions:

- **Assumption 1 (A1): Positive Definite Restricted Hessian.** $\nabla_S^2 \mathcal{L}_n(\theta^*) \succ \lambda_{\min} I$
- **Assumption 2 (A2): Irrepresentability.** There exists some $\psi \in (0, 1]$ such that
  \[
  \|\nabla_S^2 \mathcal{L}_n(\theta^*) (\nabla_S^2 \mathcal{L}_n(\theta^*))^{-1}\|_\infty \leq 1 - \psi
  \]
- **Assumption 3 (A3). Unique Minimizer.** When restricted to the true support, the solution to the $\ell_1$ penalized loss minimization problem is unique, which we denote by:
  \[
  \hat{\theta}_{\lambda_n} = \underset{\theta}{\text{argmin}} \mathcal{L}_n(\theta) + \lambda_n |\theta|_1
  \] (8)

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 $\mathcal{M}(S)$.

**Definition 2.** $\mathcal{L}_n$ is $(\alpha, \beta, \gamma)$ restricted locally separable around $\theta^*$ over the subspace $\mathcal{M}(S)$ if the remainder term $R(\Delta; \theta^*)$ satisfies:

\[
\|R(\Delta; \theta^*)\|_\infty \leq \frac{1}{\beta} \|\Delta\|_\infty^\alpha \ \forall \Delta \in B_\infty(\gamma) \cap \mathcal{M}(S)
\]

We present our main deterministic result in high dimensions.

**Theorem 2.** Let $\mathcal{L}_n$ be a $(\alpha, \beta, \gamma)$ locally separable function around $\theta^*$. If $(\lambda_n, \nabla \mathcal{L}_n(\theta^*))$ are such that,

- $\frac{\psi}{\beta} \lambda_n \geq \|\nabla \mathcal{L}_n(\theta^*)\|_\infty$.
- $\|\nabla \mathcal{L}_n(\theta^*)\|_\infty + \lambda_n \leq \min \left\{ \frac{\gamma}{2\pi}, \left(\frac{1}{2\pi}\right)^\frac{\gamma}{\gamma+2} \beta \pi^{-1} \right\}$

Then we have that $\text{support}(\hat{\theta}_{\lambda_n}) \subseteq \text{support}(\theta^*)$ and

\[
\|\hat{\theta}_{\lambda_n} - \theta^*\|_\infty \leq 2\kappa (\|\nabla \mathcal{L}_n(\theta^*)\|_\infty + \lambda_n)
\]

where $\kappa = \|\nabla_S^2 \mathcal{L}_n(\theta^*)^{-1}\|_\infty$.

\[
\hat{\theta}_{\lambda_n} = \text{argmin}_\theta \mathcal{L}_n(\theta) + \lambda_n |\theta|_1
\]
Proof. (Proof Sketch). The proof invokes the primal-dual witness argument [13] which when combined with Assumption 1-3, gives $\theta_{\lambda_n} \in M(S)$ and that $\hat{\theta}_{\lambda_n}$ is the unique solution of the restricted problem. The rest of the proof proceeds similar to Theorem 1, by constructing a suitable function $F : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}$ for which $\hat{X} = \hat{\theta}_{\lambda_n} - \theta^*$ is the unique fixed point, and showing that $F$ is contractive over $E_{\infty}(r; \theta^*)$ for $r = 2 \kappa ([|\nabla L_n(\theta^*)|]_{\infty} + \lambda_n).$ 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 $\ell(\theta, z)$ to establish $l_{\infty}$ convergence. The first step is to calculate the separability constants $(\alpha, \beta, \gamma)$ for the corresponding empirical loss function $L_n$. Next, since the loss $\ell$ is Fisher consistent, so that $\nabla L(\theta^*) = 0$, the upper bound on $|\nabla L_n(\theta^*)|_{\infty}$ can be shown to hold by analyzing the concentration of $\nabla L_n(\theta^*)$ around its mean. We emphasize that we do not impose any restrictions on the values of $(\alpha, \beta, \gamma)$. 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 $\gamma$ or $\beta$ get for any given loss $\ell$, the larger the required number of samples.

4 $\ell_{\infty}$-rates for Generative and Discriminative Model Estimation

In this section we study the $\ell_{\infty}$ 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 $\ell_1$-regularized logistic regression with the sufficient statistic as features to estimate the differential parameter. In addition to A1-A3, we assume column normalization of the sufficient statistics, i.e. $\sum_{i=1}^{n} [\nabla L_\theta(i)]^2 \leq n$. Let $\gamma_n = \max_i [\nabla L_\theta(i)]_{\infty}$, $\nu_n = \max_i |[\nabla L_\theta(i)]_2|$. Firstly, we characterize the separability of the logistic loss.

Lemma 1. The logistic regression negative log-likelihood $L_{\text{logistic}}$ from (7) is $\left(\frac{2}{\sqrt{\gamma_n \nu_n}}, \infty\right)$ restricted local separable around $\theta^*$.

Combining Lemma 1 with Theorem 2, we get the following corollary.

Corollary 3. (Logistic Regression) Consider the model in (1), then there exist universal positive constants $C_1, C_2$ and $C_3$ such that for $n \geq C_1 \kappa^2 s^2 \gamma_n^2 \nu_n^2 \log p$ and $\lambda_n = C_2 \sqrt{\frac{\log p}{n}}$, the discriminative differential estimate $\hat{\theta}_{\text{diff}}$ satisfies

$$\text{support}(\hat{\theta}_{\text{diff}}) \subseteq \text{support}(\theta_{\text{diff}}) \quad \text{and} \quad \|\hat{\theta}_{\text{diff}} - \theta_{\text{diff}}\|_{\infty} \leq C_3 \sqrt{\frac{\log p}{n}}.$$

4.2 Generative Estimation

We characterize the separability of Generative Exponential Families. The negative log-likelihood function can be written as:

$$L_n(\theta) = A(\theta) - \langle \theta, \phi_n \rangle,$$

where $\phi_n = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$. In this setting, the remainder term is independent of the data and can be written as $R(\theta) = \nabla A(\theta^* + \Delta) - \nabla A(\theta^*) - \nabla^2 A(\theta^*) \Delta$ and $\nabla L_n(\theta^*) = E[\phi(x)] - \frac{1}{n} \hat{\phi}(x)$. Hence, $[|\nabla L_n(\theta^*)|]_{\infty}$ 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 $L_{\text{IG}}$ from (4) is $(\cdot, \infty, \infty)$ locally separable around $\theta^*$.

Lemma 3. The Gaussian MRF negative log-likelihood $L_{\text{GGM}}$ from (5) is $\left(\frac{2}{3M_{\infty}^3}, \frac{2}{3M_{\infty}^3}, \frac{1}{3M_{\infty}^3}\right)$ restricted locally separable around $\Theta^*$. 
The bounds for the discriminative approach are relatively straightforward. Corollary 3 gives bounds on the $\ell_\infty$ error and establishes that $\text{support}(\hat{\theta}) \subseteq \text{support}(\theta^*)$. Since the true model parameter is $s$-sparse, $\|\theta^*\|_0 \leq s$, the $\ell_2$ error can be simply bounded as $\sqrt{s} \|\hat{\theta} - \theta^*\|_\infty$.
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 $\ell_2$ rates for differential parameter estimation at a much lower sample complexity. Unfortunately, it does not have support consistency. Hence a naïve generative estimator will have an $\ell_2$ error scaling with $\sqrt{\frac{p \log p}{n}}$, which in high dimensions, would make it unappealing. However, one can exploit the sparsity of $\theta^\ast_{\text{diff}}$ and get better rates of convergence in $\ell_2$-norm by simply soft-thresholding the generative estimate. Moreover, soft-thresholding also leads to support consistency.

**Definition 3.** We denote the soft-thresholding operator $ST_{\lambda_n} (\cdot)$, defined as:

$$ST_{\lambda_n} (\theta) = \arg\min_w \frac{1}{2} \|w - \theta\|_2^2 + \lambda_n \|w\|_1.$$ 

**Lemma 4.** Suppose $\theta = \theta^\ast + \epsilon$ for some $s$-sparse $\theta^\ast$. Then there exists a universal constant $C_1$ such that for $\lambda_n \geq 2 \|\epsilon\|_\infty$,

$$\|ST_{\lambda_n}(\theta) - \theta^\ast\|_2 \leq C_1 \sqrt{s} \|\epsilon\|_\infty \quad \text{and} \quad \|ST_{\lambda_n}(\theta) - \theta^\ast\|_1 \leq C_1 s \|\epsilon\|_\infty$$

(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 $\hat{\theta}_{\text{diff}}$ using the generative model estimates, and (b) soft-threshold the generative estimate with $\lambda_n = c \|\hat{\theta}_{\text{diff}} - \theta^\ast_{\text{diff}}\|_\infty$. An elementary application of Lemma 4 can then be shown to provide $\ell_2$ error bounds for $\hat{\theta}_{\text{diff}}$ given its $\ell_\infty$ error bounds, and that the true parameter $\theta^\ast_{\text{diff}}$ is $s$-sparse. We instantiate these $\ell_2$-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 $C_1, C_2, C_3$ such that if the number of samples scale as $n \geq C_1 \log p$, then with probability at least $1 - 1/p^{C_2}$, the soft-thresholded generative estimate of the differential parameter $ST_{\lambda_n}(\hat{\theta}_{\text{diff}})$, with the soft-thresholding parameter set as $\lambda_n = c \sqrt{\frac{\log p}{n}}$ for some constant $c$, satisfies:

$$\|ST_{\lambda_n}(\hat{\theta}_{\text{diff}}) - \theta^\ast_{\text{diff}}\|_2 \leq C_3 \sqrt{s \frac{\log p}{n}}.$$

**Lemma 6.** (Gaussian MRF) Consider the model in Equation 3, and suppose that the covariates $X_k/\sqrt{\sum_i \sigma_i^2}$ are subgaussian with parameter $\sigma^2$. Then there exist universal positive constants $C_2, C_3, C_4$ such that if the number of samples for the two generative models scale as $n_i \geq C_2 n_i^{1/2} \log p$, for $i \in \{0, 1\}$, then with probability at least $1 - 1/p^{C_4}$, the soft-thresholded generative estimate of the differential parameter, $ST_{\lambda_n}(\hat{\theta}_{\text{diff}})$, with the soft-thresholding parameter set as $\lambda_n = c \sqrt{\frac{\log p}{n}}$ for some constant $c$, satisfies:

$$\|ST_{\lambda_n}(\hat{\theta}_{\text{diff}}) - \theta^\ast_{\text{diff}}\|_2 \leq C_4 \sqrt{s \frac{\log p}{n}}.$$

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 $\ell_2$-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 $\ell_2$-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 $X|Y \sim \mathcal{N}(\mu_Y, I_p)$, where $\mu_0, \mu_1 \in \mathbb{R}^p$ are unknown
and $\mu_1 - \mu_0$ is $s$-sparse. Here, we are interested in a classifier $C: \mathbb{R}^p \to \{0, 1\}$ that achieves low classification error $\mathbb{E}_{X,Y} [1 \{C(X) \neq Y\}]$. Under this setting, it can be shown that the Bayes classifier, that achieves the lowest possible classification error, is given by the linear discriminant classifier $C^*(x) = 1 \{x^T w^* + b^* > 0\}$, where $w^* = (\mu_1 - \mu_0)$ and $b^* = \frac{n_0^2 \mu_0 - n_1^2 \mu_1}{2}$. 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 $\ell_2$ error of the estimates. Under some mild assumptions, Li et al. [3] showed that for any linear classifier $\hat{C}(x) = 1 \{x^T \hat{w} + \hat{b} > 0\}$, the excess classification error can be bounded as:

$$\mathcal{E}(\hat{C}) \leq C_1 \left( \|\hat{w} - w^*\|_2^2 + \|\hat{b} - b^*\|_2^2 \right),$$

for some constant $C_1 > 0$, and where $\mathcal{E}(C) = \mathbb{E}_{X,Y} [1 \{C(X) \neq Y\}] - \mathbb{E}_{X,Y} [1 \{C^*(X) \neq Y\}]$ is the excess 0-1 error. In other words, the excess classification error is bounded by a constant times the $\ell_2$ 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 $\hat{w}$ as well as the constant bias term $\hat{b}$ can be simply obtained via $\ell_1$-regularized logistic regression. For the generative estimate, we use our two stage estimator from Section 5, which proceeds by estimating $\hat{\mu}_0, \hat{\mu}_1$ using the empirical means, and then estimating the differential parameter by soft-thresholding the difference of the generative model parameter estimates $\hat{w}_T = \text{ST}_{\lambda_n}(\hat{\mu}_1 - \hat{\mu}_0)$ where $\lambda_n = C_1 \sqrt{\frac{\log p}{n}}$ for some constant $C_1$. The corresponding estimate for $b^*$ is given by $\hat{b}_T = -\frac{1}{2} \langle \hat{w}_T, \hat{\mu}_1 + \hat{\mu}_0 \rangle$.

**Experimental Setup.** For our experimental setup, we consider isotropic Gaussian models with means $\mu_0 = 1_p - \frac{1}{\sqrt{s}} \begin{bmatrix} 1_s \\ 0_{p-s} \end{bmatrix}$, $\mu_1 = 1_p + \frac{1}{\sqrt{s}} \begin{bmatrix} 1_s \\ 0_{p-s} \end{bmatrix}$, and vary the sparsity level $s$. For both methods, we set the regularization parameter\(^2\) as $\lambda_n = \sqrt{\log(p)/n}$. 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.
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.

References


A Proof of Theorem 1

Proof. The proof involves two steps:

- Constructing a suitable function $F : \mathbb{R}^p \mapsto \mathbb{R}^p$ for which the $\hat{\theta}$ is the unique fixed point.(See Lemma 7)
- Showing that the function is contractive over an $\ell_\infty$ ball of radius $2\kappa \|\nabla L(\theta^*)\|_\infty$.

Lemma 7. Let $\hat{\theta} = \text{argmin}_\theta L_n(\theta)$. Then $\hat{\Delta} = \hat{\theta} - \theta^*$ is the unique fixed point of $F : \mathbb{R}^p \mapsto \mathbb{R}^p$, $F(\Delta) = -\nabla^2 L_n(\theta^*)^{-1} R(\Delta; \theta^*) - \nabla^2 L_n(\theta^*)^{-1} \nabla L_n(\theta^*)$

Proof. By first order optimality of $\hat{\theta}$, we know that $\nabla L_n(\hat{\theta}) = 0$. Using invertibility of $\nabla^2 L_n(\theta^*)$, we know that for $F(\Delta) = \Delta - \nabla^2 L_n(\theta^*)^{-1} (\nabla L_n(\theta^* + \Delta))$, $\hat{\Delta}$ is the unique fixed point. Now, we can simplify $F$ as:

$$F(\Delta) = \Delta - \nabla^2 L_n(\theta^*)^{-1} (\nabla L_n(\theta^* + \Delta))$$

$$= \Delta - \nabla^2 L_n(\theta^*)^{-1} (\nabla L_n(\theta^*) - \nabla L_n(\theta^*) + \nabla L_n(\theta^*))$$

$$= \Delta - \nabla^2 L_n(\theta^*)^{-1} (R(\Delta; \theta^*) + \nabla^2 L_n(\theta^*) \Delta + \nabla L_n(\theta^*))$$

$$= -\nabla^2 L_n(\theta^*)^{-1} R(\Delta; \theta^*) - \nabla^2 L_n(\theta^*)^{-1} \nabla L_n(\theta^*)$$

where by definition, we have that $R(\Delta; \theta^*) = \nabla L_n(\theta^* + \Delta) - \nabla L_n(\theta^*) - \nabla^2 L_n(\theta^*)$.

Now to show contraction of $F$. Let $r = 2\kappa \|\nabla L_n(\theta^*)\|_\infty$ and let $\Delta \in B_\infty(r)$

$$\|F(\Delta)\|_\infty = \|-(\nabla^2 L_n(\theta^*)^{-1} R(\Delta; \theta^*) - \nabla^2 L_n(\theta^*)^{-1} \nabla L_n(\theta^*))\|_\infty$$

$$\leq \kappa \|R(\Delta; \theta^*)\|_\infty + \kappa \|\nabla L_n(\theta^*)\|_\infty$$

By our assumption, $r = 2\kappa \|\nabla L_n(\theta^*)\|_\infty \leq \gamma$, so we can upper bound $R(\Delta; \theta^*)$ using the separability of $L_n$.

$$\leq \kappa \frac{1}{\beta} (2\kappa \|\nabla L_n(\theta^*)\|_\infty)^\alpha + r/2$$

$$= \frac{\kappa}{\beta} (2\kappa)^\alpha \|\nabla L_n(\theta^*)\|_\infty^\alpha \times \frac{\kappa \|\nabla L_n(\theta^*)\|_\infty}{\kappa \|\nabla L_n(\theta^*)\|_\infty} + r/2$$

$$= \frac{1}{\beta} (2\kappa)^\alpha \|\nabla L_n(\theta^*)\|_\infty^{\alpha-1} \times \kappa \|\nabla L_n(\theta^*)\|_\infty + r/2$$

$$\leq \kappa \|\nabla L_n(\theta^*)\|_\infty + r/2$$

$$\leq r$$

where the last step follows from our assumption that $\|\nabla L_n(\theta^*)\|_\infty \leq \left(\frac{1}{2\kappa}\right)^{\frac{1}{\alpha-1}} \beta^{\frac{1}{\alpha}} r$.

Hence, we’ve shown that $F(B_\infty(r)) \subseteq B_\infty(r)$ for $r = 2\kappa \|\nabla L_n(\theta^*)\|_\infty$.

Since $F$ is continuous and $\ell_\infty$-ball is convex and compact, the contraction property coupled with Brouwer’s fixed point theorem [9], shows that there exists some fixed point $\Delta$ of $F$, such that $|\Delta|_\infty \leq 2\kappa \|\nabla L_n\|_\infty$. Lemma 7 established that $\hat{\Delta} = \hat{\theta} - \theta^*$ as the unique fixed point, hence, $|\hat{\Delta}|_\infty \leq 2\kappa \|\nabla L_n\|_\infty$.

B Proof of Theorem 2

The proof involves three steps:

- Firstly, in Lemma 8, using the primal-dual witness argument of Wainwright [13], under Assumptions 1-3, we establish that the minimizer of the unconstrained problem (8) is unique and is equal to that of the restricted problem (9), $\tilde{\theta}_{\lambda_n} = \tilde{\theta}_{\lambda_n}$.
• Using our assumption that the minimizer of the restricted function is unique, if we take partial derivatives of the Lagrangian of the restricted problem (9) with respect to the unconstrained elements, these partial derivatives must vanish at the optimum, meaning that we have the zero-gradient condition

\[ G(\theta_S) = [\nabla \mathcal{L}_n(\theta)]_S + \lambda \tilde{Z}_S = 0 \]  

(11)

where \( \theta \in \mathbb{R}^p \) is such that, it’s entries in \( S \) is equal to \( \theta_S \) and 0 in \( S^c \). The zero-gradient condition is necessary and sufficient for an optimum of the Lagrangian problem, and has a unique solution, \( [\tilde{\theta}_{\lambda_n}]_S \).

\[ G(\theta_S) = 0 \iff \theta_S = [\tilde{\theta}_{\lambda_n}]_S \]

Using this and following the proof of Theorem A, we construct a function \( F : \mathbb{R}^{|S|} \rightarrow \mathbb{R}^{|S|}, \)

\[ F(\Delta_S) = \Delta_S - (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} (G(\theta^*_S + \Delta_S)) \]

(12)

Since \( \nabla^2_{SS} \mathcal{L}_n(\theta^*) \) is invertible, \([\Delta]_S = [\tilde{\theta}_{\lambda_n} - \theta^*]_S \) is the unique optimal. \( F \) can be simplified and written as:

\[ F(\Delta_S) = \Delta_S - (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} [R(\Delta) + \nabla^2 \mathcal{L}(\theta^*) \Delta + \nabla \mathcal{L}_n(\theta^*)]_S - (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} \lambda_n \tilde{Z}_S \]

where \( \Delta \in \mathbb{R}^p \) is such that, it’s entries in \( S \) are equal to \( \Delta_S \) and 0 in \( S^c \). Following some algebra, we can show that:

\[ F(\Delta_S) = -(\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} [R(\Delta)]_S - (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} [\nabla \mathcal{L}_n(\theta^*)]_S - \lambda_n (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} \tilde{Z}_S \]

• Note that \( F(\cdot) \) is continuous. We’ll show that \( F \) is contractive over \( B_{\infty}(r) \) for \( r = 2\kappa([\nabla \mathcal{L}_n(\theta^*)]_\infty + \lambda_n) \). Let \( \Delta_S \in B_{\infty}(r) \). Then, the corresponding \( \Delta \in \mathbb{R}^p \) is such that, it’s entries in \( S \) are equal to \( \Delta_S \) and 0 in \( S^c \). Note that this \( \Delta \in B_{\infty}(r) \cap \mathcal{M}(S) \), and by assumption \( r = 2\kappa([\nabla \mathcal{L}_n(\theta^*)]_\infty + \lambda_n) \leq \gamma \), hence we can upper bound \( \|R(\Delta)\|_\infty \) using that \( \mathcal{L}_n \) is restricted locally separable around \( \theta^* \) for this radius \( r \).

\[ \|F(\Delta_S)\|_\infty \leq \|-(\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} [R(\Delta)]_S - (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} [\nabla \mathcal{L}_n(\theta^*)]_S - \lambda_n (\nabla^2_{SS} \mathcal{L}_n(\theta^*))^{-1} \tilde{Z}_S\|_\infty \]

\[ \leq \kappa \|R(\Delta)_S\|_\infty + \kappa (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n) \]

(13)

\[ \leq \kappa \|R(\Delta)_S\|_\infty + \kappa (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n) \]

(14)

\[ \leq \kappa \frac{\|R(\Delta)_S\|_\infty}{\|\Delta\|_\infty} + \kappa (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n) \]

(15)

\[ \leq \frac{\kappa}{\|\Delta\|_\infty} + \kappa (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n) \]

\[ \leq \frac{\kappa}{\beta} (2\kappa)^\alpha (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n) + r/2 \]

(16)

\[ \leq \frac{1}{\beta} (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n)^\alpha + r/2 \]

(17)

\[ \leq \frac{1}{\beta} (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n)^\alpha + r/2 \]

(18)

\[ \leq r \]

(19)

where we’ve used that by assumption, \( (\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty + \lambda_n)^\alpha \leq \frac{1}{\beta} \|
abla \mathcal{L}_n(\theta^*)_S\|_\infty \).}

**Lemma 8.** Let \( \frac{\alpha \lambda_n}{8} \geq \max\{\|\nabla \mathcal{L}_n(\theta^*)_S\|_\infty, \|R(\Delta)\|_\infty\} \). Then, \( \tilde{\theta}_{\lambda_n} = \tilde{\theta}_{\lambda_n} \) is the unique minimizer of Equation 8, i.e. \( \text{Support}(\tilde{\theta}_{\lambda_n}) \subseteq \text{Support}(\theta^*) \).

**Proof.** The proof follows from the primal-dual witness argument of [13, 11]. We provide it here completeness. The steps are outlined as:

1. By assumption we have that \( \tilde{\theta}_{\lambda_n} \) is the unique minimizer of the restricted problem.

\[ \tilde{\theta}_{\lambda_n} = \arg\min_{\theta \in \mathcal{M}(S)} \mathcal{L}_n(\theta) + \lambda_n \|\theta\|_1 \]

(20)

2. Choose \( \tilde{Z} \) such that \( \tilde{Z}_S \) is sub-differential of \( \|\tilde{\theta}_{\lambda_n}\|_1 \).

3. Choose \( \tilde{Z}_S^c \) such that \( \nabla \mathcal{L}_n(\tilde{\theta}_{\lambda_n}) + \lambda \tilde{Z} = 0 \), which ensures that \( (\tilde{\theta}_{\lambda_n}, \tilde{Z}) \) satisfy the optimality condition for (unconstrained) problem (8).
4. Verify the strict dual feasibility condition
\[ \| \tilde{Z}_{S^c} \|_\infty < 1 \]

Steps 1-3 ensure that \((\hat{\theta}_{\lambda_n}, \tilde{Z})\) satisfy the optimality conditions of (8). By construction, step 2 ensures that \(\tilde{Z}_S\) satisfies the sub-differential conditions. Then, Step 4 is needed to ensure that the remaining elements of \(\tilde{Z}\) satisfy the sub-differential conditions. If steps 1-4 succeed, then, it acts as a witness that the solution \(\hat{\theta}_{\lambda_n}\) to restricted problem is equal to solution of the unrestricted(original) problem.(8).

We now show the uniqueness of \(\hat{\theta}_{\lambda_n}\). By following an argument similar to Lemma 1 in [10],(Lemma 1(b) in [13]), any minimizer \(\hat{\theta}_{\lambda_n}\) of the original optimization problem satisfies \([\hat{\theta}_{\lambda_n}]_{S^c} = 0\). Thus, since \(\hat{\theta}_{\lambda_n}\) is the only optimal vector for the restricted optimization problem(by assumption), we conclude that \(\hat{\theta}_{\lambda_n} = \hat{\theta}_{\lambda_n}\).

Now, to show that \(\| \tilde{Z}_{S^c} \|_\infty < \). Let \(\hat{\Delta} = \hat{\theta}_{\lambda_n} - \theta^\ast\). So, we know \(\hat{\Delta}_{S^c} = 0\). Using this we can write the gradient condition. For brevity, let \(\hat{\theta} = \hat{\theta}_{\lambda_n} \), \(\nabla^2 L_n(\theta^\ast) = \Gamma\).

\[
\nabla L_n(\hat{\theta}) + \lambda_n \tilde{Z} = 0
\]

(21)

We can rewrite the gradient:
\[
\Gamma_{SS} \hat{\Delta} + [R(\hat{\Delta})]_S + (\nabla L_n(\theta^\ast))_S + \lambda_n \tilde{Z}_S = 0
\]

(22)

\[
\Gamma_{S^cS} \hat{\Delta} + [R(\hat{\Delta})]_{S^c} + (\nabla L_n(\theta^\ast))_{S^c} + \lambda_n \tilde{Z}_{S^c} = 0
\]

(23)

Since, \(\Gamma_{SS}\) is invertible,

\[
\hat{\Delta} = \Gamma_{SS}^{-1} \left[ [R(\hat{\Delta})]_S + (\nabla L_n(\theta^\ast))_S + \lambda_n \tilde{Z}_S \right]
\]

(24)

Plugging this into \(\tilde{Z}_{S^c}\),
\[
\| \tilde{Z}_{S^c} \|_\infty = \frac{1}{\lambda_n} \left\| [R(\hat{\Delta})]_{S^c} + (\nabla L_n(\theta^\ast))_{S^c} + \Gamma_{S^cS}(\Gamma_{SS})^{-1} \left[ [R(\hat{\Delta})]_S + (\nabla L_n(\theta^\ast))_S + \lambda_n \tilde{Z}_S \right] \right\|_\infty
\]

(25)
\[
\leq \frac{2 - \psi}{\lambda_n} \left( \left\| [R(\hat{\Delta})]_S \right\|_\infty + \left\| (\nabla L_n(\theta^\ast))_S \right\|_\infty \right) + (1 - \psi)
\]

(26)
\[
\leq \frac{2 - \psi}{\lambda_n} \psi \lambda_n / 4 + (1 - \psi)
\]

(27)
\[
\leq 1 - \psi / 2 < 1
\]

(28)

where we’ve used that \(\| \Gamma_{S^cS}(\Gamma_{SS})^{-1} \|_\infty \leq 1 - \psi\), and \(\| \tilde{Z}_S \|_\infty \leq 1\) in the first step. The next step follows from our assumption that \(\max\{ \left\| \nabla L_n(\theta^\ast) \right\|_\infty, \left\| R(\Delta) \right\|_\infty \} \leq \psi \lambda / 8\). \(\square\)

C Proof of Lemma 1

**Proof.** The proof follows the analysis of [10]. Let \(L_n\) be the logistic loss. Let \(\Delta \in \mathcal{B}_\infty(\gamma) \cap \mathcal{M}(S)\). For the logistic loss, the remainder term can be written as:

\[
R(\Delta) = \nabla L_n(\theta^\ast + \Delta) - \nabla L_n(\theta^\ast) - \nabla^2 L_n(\theta^\ast) \Delta
\]

(29)

\[
= [\nabla^2 L_n(\theta^\ast) - \nabla^2 L_n(\theta^\ast)] \Delta
\]

(30)
where from mean-value theorem \( \bar{\theta} = t(\theta^* + \Delta) + (1 - t)\theta^* \) is point on line joining \( \theta^* + \Delta \) and \( \theta^* \). Now, looking at the \( j \)th entry, we get:

\[
[R(\Delta)]_j = [\nabla^2 L_n(\bar{\theta}) - \nabla^2 L_n(\theta^*)]_j \Delta 
\]

\[
= \left[ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T (\eta(\bar{\theta}^T x_i) - \eta(\theta^* T x_i)) \right]_j \Delta 
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (\eta(\bar{\theta}^T x_i) - \eta(\theta^* T x_i)) x_i^{(j)} x_i^T \Delta 
\]

where \( \eta(t) = \frac{\exp(t)}{(1 + \exp(t))^2} \). Again applying, mean-value theorem, where \( \bar{\theta} \) is another point between \( \theta^* + \Delta \) and \( \theta^* \), we get that

\[
[R(\Delta)]_j = \frac{1}{n} \sum_{i=1}^{n} \eta'(\bar{\theta}^T x_i)(\bar{\theta} - \theta^*)x_i^{(j)} x_i^T \Delta 
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left[ \eta'(\bar{\theta}^T x_i)x_i^{(j)} \right]_{a_i} \left[ (\bar{\theta} - \theta^*) T x_i x_i^T \Delta \right]_{b_i} 
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} a_i b_i 
\]

By holder’s inequality

\[
|[R(\Delta)]_j| = \frac{1}{n} \left| \sum_{i=1}^{n} a_i b_i \right| \leq \frac{1}{n} \|a\|_\infty \|b\|_1 
\]

Now \( |a|_\infty \leq \gamma_n \). Also, observe that by definition \( \bar{\theta} - \theta^* = t \Delta \)

\[
\frac{1}{n} \|b\|_1 = t \Delta^T \left\{ \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T \right\} \Delta 
\]

\[
= t \Delta^T S \left\{ \frac{1}{n} \sum_{i=1}^{n} (x_i s(x_i))^T \right\} S \Delta 
\]

\[
\leq \nu^2 \|\Delta S\|_2^2 
\]

\[
\leq \nu^2 \|\Delta\|_\infty^2 
\]

where we’ve used that, \( \Delta \in \mathcal{M}(S) \). Combining the above, we get that,

\[
|R(\Delta)|_\infty \leq \gamma_n \nu^2 s \|\Delta\|_\infty^2 
\]

Hence, the logistic loss is \( (2, \frac{1}{\gamma_n \nu^2 s}, \infty) \) separable around \( \theta^* \).

### D Proof of Corollary 3

In Lemma 1, we’ve already established the separability of logistic loss. Assume \( \lambda_n = c \cdot \|\nabla L_n(\theta^*)\|_\infty \). We need to control \( \|\nabla L_{\text{logistic}}(\theta^*)\|_\infty \), so that \( \|\nabla L_n(\theta^*)\|_\infty \leq \min \left\{ \frac{\gamma_n}{\alpha}, \left( \frac{1}{2n} \right)^{\alpha/\beta} \right\} \).

**Proof.**

\[
L_{\text{logistic}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[ -y_i (\theta^T \phi(x_i)) + \log(1 + \exp(\theta, \phi(x_i))) \right] 
\]

\[
L_{\text{logistic}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i) \left[ -y_i + u(\theta^T \phi(x_i)) \right], 
\]
where \( u(t) = \frac{1}{1 + \exp(-t)} \). Note that \( u(\theta^T \phi(x_i)) = \mathbb{P}(Y_i = 1|\phi(x_i)) = \mathbb{E}(Y_i) \).

The jth-coordinate of \( L_{\logistic}(\theta^*) \) can then be written as:

\[
[\nabla L_{\logistic}(\theta^*)]_j = -\frac{1}{n} \sum_{i=1}^{n} [\phi(x_i)]_j (y_i - \mathbb{E}[Y_i])
\]  

(44)

By Assumption, we have that \( \sum_{i=1}^{n} [\phi(x_i)]_j^2 \leq n \), therefore the random variable \( -\frac{1}{n} \sum_{i=1}^{n} [\phi(x_i)]_j y_i \) is subgaussian\( (\sigma^2 = 1/n^2 * (\sum_{i=1}^{n} [\phi(x_i)]_j^2) * (1/4)) = 1/(4n) \).

Hence, we have each co-ordinate of \( \nabla L_{\logistic}(\theta^*) \) is subgaussian\( (\sigma^2 = \frac{1}{4n}) \). By union bound and subgaussian tail bounds, we have:

\[
\mathbb{P}(\|\nabla L_{\logistic}(\theta^*)\|_\infty \geq \tau) \lesssim p^2 \exp(-\tau^2 n c_0)
\]

(45)

for some constant \( c_0 \). To get a high probability bound, we essentially want \( p^2 \exp(-\tau^2 n c_0) < 1/p^c \) for some \( c > 0 \). We get that \( n >> \frac{1}{\tau^2} \log p \) suffices.

From Lemma 1, we know that \( \alpha = 2 \), \( \beta = \frac{1}{\sqrt{\gamma_n \nu_n^2}} \), \( \gamma = \infty \), so, we need to put \( \tau = \beta = \frac{1}{\sqrt{\gamma_n \nu_n^2}} \). Hence, we get that \( n >> \frac{2 s^2 \gamma_n \nu_n^2 \log p}{\log p} \) samples are enough to ensure that \( |\nabla L_n(\theta^*)|_\infty \leq \min \left\{ \frac{2}{\sqrt{n}}, (\frac{1}{2\nu_n}) \sqrt{\frac{\tau}{\gamma_n \nu_n}} \right\} \).

Now, putting \( \tau = c\sqrt{\frac{\log p \ell}{n}} \) in (45) we get that \( \|\nabla L_{\logistic}(\theta^*)\|_\infty \lesssim \sqrt{\frac{\log p \ell}{n}} \) with high probability, whenever \( n \gtrsim \log p \). Since, we put \( \lambda_n = c \|\nabla L_{\logistic}(\theta^*)\|_\infty \), we get that \( \|\nabla L_{\logistic}(\theta^*)\|_\infty + \lambda_n \lesssim \sqrt{\frac{\log p \ell}{n}} \).

\[\] E Proof of Lemma 2

**Proof.** From Equation 4, we see that

\[ R(\Delta) = (\theta^* + \Delta) - \theta^* - \Delta = 0 \]

So, we can choose \( \beta = \gamma = \infty \).

\[\]

F Proof of Corollary 4

**Proof.** From Lemma 2, we know that \( \beta = \gamma = \infty \). Hence, there is no initial sample complexity. We know that \( \nabla L_n(\theta^*) = \theta^* - \frac{1}{n} \sum_{i=1}^{n} x_i \). Each co-ordinate of \( [\nabla L_n(\theta^*)]_j \sim \mathcal{N}(0, 1/n) \). By known concentration results on maximum of \( p \) gaussians, we know that when \( n >> \log p \), then \( |\nabla L_n(\theta^*)|_{\infty} \lesssim \sqrt{\frac{\log p \ell}{n}} \) with high probability. Hence, for individual parameters \( \hat{\theta}_1 \) and \( \hat{\theta}_0 \) we have \( \ell_{\infty} \)-error of \( O(\sqrt{\frac{\log p \ell}{n}}) \), hence, differential estimate also has \( \ell_{\infty} \)-error of \( O(\sqrt{\frac{\log p \ell}{n}}) \).

\[\]

G Proof of Lemma 3

**Proof.** The proof is based on Lemma 5 by [11] and follows from matrix expansion techniques. We provide it’s key steps for the sake of completeness. Let \( \Delta \in B_{\infty}(\gamma) \cap M(S) \). For \( L_{\logistic} \), the remainder can be written as:

\[
R(\Delta) = (\theta^* + \Delta)^{-1} - \theta^*-1 + \theta^*-1 \Delta \theta^*-1
\]

(46)

\[
= (I + \theta^*-1 \Delta)^{-1} (\theta^*)^{-1} - \theta^*-1 + \theta^*-1 \Delta \theta^*-1
\]

(47)

\[
= \sum_{k=0}^{\infty} (-1)^k ((\theta^*)^{-1} \Delta)^k (\theta^*)^{-1} - \theta^*-1 + \theta^*-1 \Delta \theta^*-1
\]

(48)

\[
= \theta^*-1 \Delta \theta^*-1 \Delta J \theta^*-1
\]

(49)
where \( J = \sum_{k=0}^{\infty} (-1)^k (\Theta^*)^{-1} \Delta)^k \).

\[
\|R(\Delta)\|_{\infty} = \max_{e_i, e_j} |e_i^T \Theta^* - 1 \Delta \Theta^* - 1 \Delta J \Theta^* - 1 e_j|
\leq \max_i \|e_i^T \Theta^* - 1 \Delta \Theta^* - 1\|_{\infty} \max_j \|\Theta^* - 1 \Delta J \Theta^* - 1 e_j\|_1
\leq \|\Delta\|_{\infty} \|((\Theta^*)^{-1})^3 \|_{\infty} \|\Delta\|_{\infty} \|J^T\|_{\infty}
\]

where \( J = \sum_{k=0}^{\infty} (-1)^k (\Theta^*)^{-1} \Delta)^k \).

\[
\|J^T\|_{\infty} \leq \sum_{k=0}^{\infty} \|\Delta(\Theta^*)^{-1}\|_{k} \leq \frac{1}{1 - d\|\Delta(\Theta^*)^{-1}\|_{\infty}} \leq \frac{1}{1 - d\|\Delta\|_{\infty} \|\Theta^* - 1\|_{\infty}}
\]

Now, by assumption we know that \( \Delta \in \mathcal{M}(S) \), hence the support pattern of \( \Delta \) is the same as that of \( \Theta^* \), which implies that \( \|\Delta\|_{\infty} \leq d\|\Delta\|_{\infty} \).

So, we have that

\[
R(\Delta) = \|((\Theta^*)^{-1})^3 d \|\Delta\|_{\infty}^{2} \frac{1}{1 - d\|\Delta(\Theta^*)^{-1}\|_{\infty}}
\]

For \( \|\Delta\|_{\infty} \leq \frac{1}{3d\kappa_\Sigma^2} \), we have that

\[
R(\Delta) \leq 1.5d\kappa_\Sigma^3 \|\Delta\|_{\infty}^2 \quad \forall \|\Delta\|_{\infty} < \frac{1}{3d\kappa_\Sigma^2}
\]

So, the loss is \((2, \frac{1}{1.5d\kappa_\Sigma^2}, \frac{1}{3d\kappa_\Sigma^2})\) separable.

\[\square\]

**H Proof of Corollary 5**

**Proof.** In Lemma 3, we’ve already established the separability of \( \mathcal{L}_{nGGM} \). Assume \( \lambda_n = c \|\nabla \mathcal{L}_{nGGM}(\Theta^*)\|_{\infty} \). We need to control \( \|\nabla \mathcal{L}_{nGGM}(\Theta^*)\|_{\infty} \), so that \( \|\nabla \mathcal{L}_{nGGM}(\Theta^*)\|_{\infty} \leq \min \left\{ \frac{2}{2\kappa}, \left(\frac{1}{2\kappa}\right)^{\frac{\alpha}{\alpha - 1}} \beta^{\frac{\alpha}{\alpha - 1}} \right\} \).

We know that \( \nabla \mathcal{L}_{nGGM}(\Theta^*) = \hat{\Sigma}_n - \Sigma^* \), where \( \hat{\Sigma}_n \) is the empirical covariance matrix. To control \( \|\nabla \mathcal{L}_{nGGM}(\Theta^*)\|_{\infty} \), we use the following result by [11].

**Lemma 9** (Lemma 1,[11]). Consider a zero-mean random vector \((X_1, \ldots, X_p)\) with covariance such that each \( X_i / \sqrt{\Sigma_{ii}} \) subGaussian with parameter \( \sigma^2 \). Given \( n \) iid samples, then the associated sample covariance \( \hat{\Sigma}_n \) satisfies the tail bound(modulo constants)

\[
P \left( \left| \hat{\Sigma}_{ij} - \Sigma_{ij}^* \right| > \delta \right) \leq 4 \left( -c n \delta^2 \right)
\]

Taking union bound over \( p^2 \) terms, we get that, for \( n >> \left\{ \max \left\{ \frac{4\kappa^2}{\sigma^2}, (2\kappa)^{2\alpha/(\alpha - 1)} \left( \frac{1}{2} \right)^{2/(\alpha - 1)} \right\} \log p, \right. \) with high probability,

\[
\|\nabla \mathcal{L}_{nGGM}(\Theta^*)\|_{\infty} = \left\| \hat{\Sigma}_n - \Sigma^* \right\|_{\infty} \leq \min \left\{ \frac{\gamma}{2\kappa}, \left( \frac{1}{2\kappa} \right)^{\frac{\alpha}{\alpha - 1}} \beta^{\frac{\alpha}{\alpha - 1}} \right\}
\]

From Lemma 3, we know that \( \alpha = 2, \beta = \frac{1}{1.5d\kappa_\Sigma^2}, \gamma = \frac{1}{2d\kappa_\Sigma^2} \). Plugging these values, we get that for \( n >> d^2 \log p \), \( \|\hat{\Theta}_1 - \Theta^*_1\|_{\infty} \leq \sqrt{\frac{\log p}{n}} \) and similarly, \( \|\hat{\Theta}_0 - \Theta_0^*\|_{\infty} \leq \sqrt{\frac{\log p}{n}} \).

Moreover from Theorem 1, we get that \( \text{support}(\hat{\Theta}_i) \subseteq \text{support}(\Theta^*_i) \) for \( i = 0, 1 \).

By triangle inequality, \( \|\hat{\Theta}_{nGGM} - \Theta^*_n\|_{\infty} \leq \sqrt{\frac{\log p}{n}} \).

16
I Proof of Lemma 4

Proof. This statement is a special instantiation of Theorem 1 in [7]. A proof of the lemma can also be found in [4](Theorem 7).

J Cross-Validated Experiments

Figure 3: Effect of sparsity on Classification Error. Observe the effect of \( s \) on the burn-in period for logistic before it starts classifying. All experiments plotted the average of 20 trials. In all experiments we set the regularization parameter \( \lambda_n \) using cross-validation.