# On the sample complexity of parameter estimation in logistic regression with normal design

**Daniel Hsu** DJHSU@CS.COLUMBIA.EDU
*Department of Computer Science, Columbia University*

**Arya Mazumdar** ARYA@UCSD.EDU
*Halıcıoğlu Data Science Institute, University of California, San Diego*

**Editors:** Shipra Agrawal and Aaron Roth

## Abstract

The logistic regression model is one of the most popular data generation model in noisy binary classification problems. In this work, we study the sample complexity of estimating the parameters of the logistic regression model up to a given $\ell_2$ error, in terms of the dimension and the inverse temperature, with standard normal covariates. The inverse temperature controls the signal-to-noise ratio of the data generation process. While both generalization bounds and asymptotic performance of the maximum-likelihood estimator for logistic regression are well-studied, the non-asymptotic sample complexity that shows the dependence on error and the inverse temperature for parameter estimation is absent from previous analyses. We show that the sample complexity curve has two change-points in terms of the inverse temperature, clearly separating the low, moderate, and high temperature regimes.

**Keywords:** logistic regression, parameter estimation, sample complexity, normal design

## 1. Introduction

This paper studies the sample complexity of estimating the parameter vector in the logistic regression model under a normal design, with particular attention paid to the dependence on the *dimension* $d$, the *inverse temperature* $\beta \geq 0$, and the *target error* $\epsilon \in (0, 1)$. We show how the form of the sample complexity changes depending on the particular relationship between the inverse temperature and the target error.

Our statistical model is as follows. The parameter space is the unit sphere $S^{d-1} = \{\theta \in \mathbb{R}^d : \|\theta\| = 1\}$, where $\|\cdot\|$ denotes the $\ell_2$ norm on $\mathbb{R}^d$. The covariate vector $\mathbf{x}$ is a $d$-dimensional standard normal random vector. Conditional on $\mathbf{x}$, the response $\mathbf{y}$ is a binary $\{-1, 1\}$-valued (Bernoulli) random variable. If the parameter vector in our model is $\theta \in S^{d-1}$, then $\mathbf{y} = 1$ with probability $g'(\beta \mathbf{x}^\mathsf{T}\theta)$, where $g'$ is the standard logistic function $g'(\eta) = 1/(1 + e^{-\eta})$, which is the derivative of the log partition function $g(\eta) = \ln(1 + e^{\eta})$. The inverse temperature $\beta$, which is the norm of the coefficient vector on $\mathbf{x}$ appearing in the mean parameter, is regarded as a model hyperparameter, and it governs the signal-to-noise ratio of this data generation process. In particular, when $\beta = 0$, the response is pure noise—a fair coin flip—with no dependence on the covariates. When $\beta = +\infty$, the response is fully determined by a homogeneous linear classifier, simply denoting the side of the hyperplane $\{x \in \mathbb{R}^d : x^\mathsf{T}\theta = 0\}$ that $\mathbf{x}$ lies on.

We assume the observed data $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$ are independent copies of $(\mathbf{x}, \mathbf{y})$, with distribution determined by the unknown parameter vector $\theta^\star \in S^{d-1}$ (and inverse temperature $\beta$).

For a given $\epsilon \in (0,1)$, the goal is to find an estimate $\hat{\theta} = \hat{\theta}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) \in S^{d-1}$ based on these data (and possibly also $\beta$) such that $\hat{\theta}$ satisfies

$$\|\hat{\theta} - \theta^\star\| \leq \epsilon$$

(either in expectation or with high probability over the realization of the data).

The *sample complexity* $n^\star(d, \beta, \epsilon)$ is the smallest sample size $n$ such that the above task is achievable by some estimator. Intuitively, the estimation error for the maximum likelihood estimator here is asymptotically normal; leading one to believe that the sample complexity must scale as $d/\epsilon^2$. However, as $\beta$ increases to $+\infty$, the estimation problem gradually becomes a noiseless halfspace learning problem. The sample complexity for the latter problem is known to be $d/\epsilon$. This means that the inverse temperature $\beta$ should play a crucial role in the sample complexity; going from small $\beta$ to large $\beta$, the sample complexity curve must have one or more change-points. In this paper, we coarsely determine this dependence on $\beta$. In particular, we show that, up to logarithmic factors in $d$ and $1/\epsilon$ in the expressions below, the sample complexity satisfies

$$n^\star(d, \beta, \epsilon) \asymp \begin{cases} \dfrac{d}{\beta^2 \epsilon^2} & \text{if } \beta \lesssim 1 \text{ (high temperatures);} \\ \dfrac{d}{\beta \epsilon^2} & \text{if } 1 \lesssim \beta \lesssim 1/\epsilon \text{ (moderate temperatures);} \\ \dfrac{d}{\epsilon} & \text{if } \beta \gtrsim 1/\epsilon \text{ (low temperatures).} \end{cases}$$

### 1.1. Motivation

Our original motivation for studying this problem comes from the application to noisy one-bit (compressive) sensing, in which only a single bit is retained per linear measurement of a signal (Boufounos and Baraniuk, 2008). In that context, Plan and Vershynin (2012) proposed a robust linear estimator that is well-behaved under a variety of observation models, including the logistic regression model that we consider. This estimator (which in the present context is essentially the same as the "Average" algorithm of Servedio (1999)) was shown by Plan et al. (2017) to have sample complexity

$$O\left(\frac{d}{\min\{\beta^2, 1\}\epsilon^2}\right),$$

improving on an earlier analysis of Plan and Vershynin (2012) that had a worse dependence on $\epsilon$; note that here we do not assume that $\theta^\star$ is sparse. The optimality of this estimator in the high-temperature regime ($\beta \lesssim 1$) is readily established as a standard application of Fano's inequality (see, e.g., Chen et al., 2016, Appendix C.1). However, it was unclear whether this sample complexity is optimal in other regimes. In particular, in the zero-temperature regime (i.e., $\beta = +\infty$), the $\mathbf{y}_i$ are determined by the sign of $\mathbf{x}_i^\mathsf{T}\theta^\star$, so the problem becomes equivalent to that of PAC learning homogeneous linear classifiers under spherically symmetric distributions on $\mathbf{x}$. For that problem, the sample complexity is $\Theta(d/\epsilon)$, as established by Long (1995, 2003); note that the dependence on $\epsilon$ is considerably reduced.[1] Jacques et al. (2013) also gives a statement of the lower bound in the context of one-bit compressive sensing.

1. Servedio (1999) studied the "Average" algorithm in this context, allowing for the possibility that each observed label is independently flipped with some fixed probability $\eta \in [0, 1/2)$, and obtains a sample complexity upper bound of $O(d/((1-2\eta)\epsilon)^2)$.

Lower bounds on the sample complexity in our setting do not directly follow from standard lower bounds from statistical learning theory for homogeneous linear classifiers (e.g., Devroye and Lugosi, 1995), as the distributions exhibiting the lower bounds generally do not conform to our statistical model. In particular, the support of $\mathbf{x}$ in these lower bounds is typically taken to be a shattered finite set of points, and the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ may not be of the form in logistic regression. The exceptions, as mentioned above, are those based on Fano's inequality for the high temperature regime, and the lower bound in the zero temperature regime of Long (1995).

Our upper and lower bounds resolve the dependence of the sample complexity on the inverse temperature (up to logarithmic factors in $1/\epsilon$ and $d$), particularly in the regime where $1 \lesssim \beta < +\infty$.

### 1.2. Techniques

**Lower bounds.** Our lower bounds for the high and moderate temperature regimes (Theorem 1) are proved using Fano's inequality; for convenience, we use a Bayesian version due to Zhang (2006), although the "Generalized Fano" approach of Han and Verdú (1994) would also work. The bound in the high temperature regime is a "textbook" application that uses a uniform quadratic bound on the Kullback–Leibler (KL) divergence between the distributions determined by nearby parameter vectors. However, the moderate temperature regime requires a more refined analysis that does not appear to be standard. To facilitate the required computation, we use the Bregman divergence form of the KL divergence between Bernoulli distributions.

For the low temperature regime, this version of Fano's inequality cannot be used, as the aforementioned KL divergence becomes unbounded. The basic form of Fano's inequality is applicable, as the conditional entropy of the (randomly chosen) parameter given the data can be estimated. However, we were not able to obtain the optimal lower bound this way in this low temperature regime. Instead, we replicate the combinatorial argument of Long (1995) with modifications to handle finite $\beta$ (Theorem 3).

**Upper bounds.** To establish upper bounds on the sample complexity, it is natural to consider the maximum likelihood estimator (MLE), which is equivalent to finding $\theta \in S^{d-1}$ that minimizes the empirical risk with respect to the logistic loss: $(1/n)\sum_{i=1}^{n} g(-\beta \mathbf{y}_i \mathbf{x}_i^\mathsf{T}\theta)$. The convexity of the logistic loss potentially makes the MLE computationally tractable (perhaps after extending the parameter space to the ball). The risk of a given $\theta$ is the expected value of this empirical risk, and the excess risk relative to that of $\theta^\star$ is the KL divergence between the distributions determined by $\theta$ and $\theta^\star$. This KL divergence can, in turn, be related to the parameter error $\|\theta - \theta^\star\|$ using our analysis from the lower bound. However, bounding the excess risk sharply enough appears to be challenging. A standard approach in statistical learning theory is to use techniques like Rademacher averages from the theory of empirical processes to relate excess risks to the excess empirical risks. Unfortunately, using such tools designed for smooth loss functions like the logistic loss (e.g., Srebro et al., 2010) leads to a sample complexity with suboptimal dependences on $\epsilon$ and $\beta$. The distribution of $(\mathbf{x}, \mathbf{y})$ satisfies the Tsybakov-Mammen margin condition $\Pr[|g'(\beta \mathbf{x}^\mathsf{T}\theta^\star) - 1/2| \le t] \le C_0 t^{\alpha/(1-\alpha)}$ with $C_0 = O(1/\beta)$ and $\alpha = 1/2$ (Mammen and Tsybakov, 1999), but this only improves excess classification error bounds, not parameter estimation error.

A different approach is to directly analyze the MLE by computing tight Taylor expansions of the estimation error, and using properties such as self-concordance of the logistic function, in a way that keeps track of the dimension dependence (e.g., He and Shao, 2000; Portnoy, 1988; Bach, 2010; Ostrovskii and Bach, 2021), and, ideally, the inverse temperature. These analyses of the MLE

match the performance guarantees in the leading order terms as predicted by classical asymptotic analysis. However, they essentially treat $\beta$ as a constant (resulting in suboptimal dependence on $\beta$ in lower order terms), which we cannot afford to do as the moderate and low temperature regimes are defined by the comparison of $\beta$ to $1/\epsilon$. Redoing the analysis entirely in our setting boils down to showing $(\theta - \theta^\star)^\mathsf{T} \nabla L(\theta) > 0$ for all $\theta$ that are $\epsilon$ away from the true parameter vector $\theta^\star$, with $L$ being the negative log-likelihood function. At this point, we need a tight approximation of $\nabla L(\theta)$ around $\nabla L(\theta^\star)$, which one may hope to obtain using the self-concordance property. However, in our setting, the inner product depends on $\theta$, and a uniform bound does not give the optimal scaling.

A very recent work of Kuchelmeister and van de Geer (2023) studies the estimation of the parameter vector in a probit regression model (under a normal design) by maximizing the likelihood under a (misspecified) logistic regression model. They directly use techniques from empirical process theory specialized to the normal design to provide upper bounds on the sample complexity needed to estimate $\theta^\star$ up to a given $\ell_2$ error (under certain assumptions on the signal-to-noise ratio in the probit model; they are also concerned with estimation of the signal-to-noise ratio itself, which is beyond the scope of our work). We leverage essential parts of their analysis to establish our sample complexity upper bound in the moderate and low temperature regimes of our problem, which requires new moment bounds in the logistic regression observation model. We also give sample complexity bounds based on empirical risk minimization (for zero-one loss) in the low temperature regime, mostly using standard techniques. As mentioned above, an optimal estimator for the high temperature regime was already known (Plan et al., 2017).

Some of estimators we study are applicable under other designs, but the analyses in this work make heavy use of symmetry properties of the normal distribution. Bounds on various Gaussian integrals essential in our proofs are collected in Appendix A.

### 1.3. Other related works

**Improper learning.** Several prior works analyze "improper learning" algorithms for (possibly misspecified) logistic regression (e.g., Kakade and Ng, 2004; Zhang, 2006; Hazan et al., 2014; Foster et al., 2018; Mourtada and Gaïffas, 2022) that do not necessarily produce an estimate of $\theta^\star$, which may not be sensible anyway if the model is misspecified. Instead, the goal of these algorithms is to achieve low prediction error guarantees. The lower bound of Hazan et al. (2014) exploits misspecification to show that larger parameter norm is more detrimental to proper online learning than it is to improper online learning. In our well-specified setting for parameter estimation, the parameter norm has a very different effect.

**High-dimensional proportional asymptotic analysis.** Another line of work considers the performance of MLE and regularized variants in the proportional asymptotic regime, where both $d$ and $n$ increase to infinity with $d/n$ tending to a constant $\delta > 0$. In this setup, Sur and Candès (2019) are able to precisely characterize the region in the plane of $\delta$ and $\beta$ where the MLE exists. Salehi et al. (2019) consider the various regularized variants of MLE and characterize their asymptotic performance. Neither work directly reveals the dependence of the sample complexity on $\beta$.

### 1.4. Notations

For notational convenience, we assume that a Bernoulli distribution $\mathrm{Bern}(p)$ has support on $\{-1, 1\}$, with the "mean parameter" $p$ being the probability of 1. We occasionally associate each $\theta \in S^{d-1}$

with a homogeneous linear classifier $h_\theta \colon \mathbb{R}^d \to \{-1, 1\}$, given by $h_\theta(x) = 1$ if $x^\mathsf{T}\theta > 0$ and $h_\theta(x) = -1$ otherwise. For any $\theta, \theta' \in S^{d-1}$, let $\mathrm{err}_\theta(\theta') = \Pr(h_{\theta'}(\mathbf{x}) \neq \mathbf{y}) = \Pr(\mathbf{y}\mathbf{x}^\mathsf{T}\theta' \leq 0)$ be the error rate of $h_{\theta'}$ when the distribution of $(\mathbf{x}, \mathbf{y})$ is specified by parameter $\theta$.

## 2. Lower bounds on the sample complexity

In this section, we give two lower bounds on the sample complexity.

### 2.1. Moderate and high temperatures

The following theorem establishes our sample complexity lower bound for moderate and high temperatures.

**Theorem 1** *Fix $\epsilon \in (0, 1)$. Suppose $\hat{\theta}$ is an estimator that, for any $\theta^\star \in S^{d-1}$,*

$$\mathbb{E}\Big[\|\hat{\theta}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) - \theta^\star\|\Big] \leq \epsilon,$$

*where the expectation is with respect to the data with distribution determined by $\theta^\star$. Then the sample size $n$ must satisfy*

$$n \geq \frac{(d-1)\ln 2 - \ln 4}{32\epsilon^2\beta\min\{\beta, 2\sqrt{2/\pi}\}}.$$

The proof of Theorem 1 uses the following information theoretic lower bound method due to Zhang (2006).

**Lemma 2 (Theorem 6.1 of Zhang, 2006)** *Let $\Pi$ be a probability measure on a parameter space $\Theta$ indexing a family of probability measures $(P_\theta)_{\theta\in\Theta}$ on a data space $\mathcal{Z}$, and let $L \colon \Theta \times \Theta \to \mathbb{R}$ be a loss function. Let $\boldsymbol{\theta} \sim \Pi$ and $\mathbf{Z} \mid \boldsymbol{\theta} \sim P_{\boldsymbol{\theta}}$. For any (possibly randomized) estimator $\hat{\theta} \colon \mathcal{Z} \to \Theta$,*

$$\mathbb{E}\Big[L(\boldsymbol{\theta}, \hat{\theta}(\mathbf{Z}))\Big] \geq \frac{1}{2}\sup\left\{\varepsilon : \inf_{\theta\in\Theta} -\ln(\Pi(B(\theta, \varepsilon))) \geq 2\kappa + \ln 4\right\}$$

*where $B(\theta, \varepsilon) = \{\theta' \in \Theta : L(\theta, \theta') < \varepsilon\}$ and $\kappa = \mathbb{E}_{(\boldsymbol{\theta},\boldsymbol{\theta}')\sim\Pi\otimes\Pi}[\mathrm{KL}(P_{\boldsymbol{\theta}}\|P_{\boldsymbol{\theta}'})]$.*

**Proof of Theorem 1.** We prove the contrapositive. Assume that

$$n < \frac{(d-1)\ln 2 - \ln 4}{32\epsilon^2\beta\min\{\beta, 2\sqrt{2/\pi}\}}.$$

Set $\delta = 4\epsilon$. Fix a unit vector $u \in S^{d-1}$, and let $C = \{\theta \in S^{d-1} : \|\theta - u\| \leq \delta\}$ be the spherical cap of radius $\delta$ around $u$. Let $\Theta = S^{d-1}$, $\Pi$ be the uniform measure on $C$, and $L(\theta, \theta') = \|\theta - \theta'\|$. For each $\theta \in \Theta$, we let $P_\theta$ denote the joint distribution of the data $\mathbf{Z} = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$ as determined by our model with parameter $\theta$. Observe that $\Pi(B(\theta, \varepsilon)) \leq (\varepsilon/\delta)^{d-1}$ for any $\varepsilon \leq \delta$. Therefore, provided that

$$(d-1)\ln 2 > 2\kappa + \ln 4,$$

we have $\mathbb{E}[\|\boldsymbol{\theta} - \hat{\theta}(\mathbf{Z})\|] > \delta/4 = \epsilon$ by Lemma 2. So it remains to establish the above displayed inequality.

Since the $n$ data are i.i.d. in each of $P_\theta$ and $P_{\theta'}$, and the marginal distribution of $\mathbf{x}$ is the same in both $P_\theta$ and $P_{\theta'}$, the chain rule for KL divergence implies

$$\mathrm{KL}(P_\theta \| P_{\theta'}) = n\,\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta)) \| \,\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta')))\big].$$

Let $\mathbf{z} = \mathbf{x}^\mathsf{T}\theta$ and $\mathbf{z}' = \mathbf{x}^\mathsf{T}\theta'$, so each of $\mathbf{z}$ and $\mathbf{z}'$ is a standard normal random variable, and the correlation $\rho$ between $\mathbf{z}$ and $\mathbf{z}'$ satisfies

$$\rho = \theta^\mathsf{T}\theta' = 1 - \frac{1}{2}\|\theta - \theta'\|^2 \geq 1 - \frac{1}{2}(2\delta)^2 = 1 - 2\delta^2,$$

where we have used the triangle inequality $\|\theta - \theta'\| \leq \|\theta - u\| + \|\theta' - u\| \leq 2\delta$. By Lemma 19 (see Appendix A),

$$\begin{aligned}
\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta)) \| \,\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta')))\big] &= \mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta \mathbf{z})) \| \,\mathrm{Bern}(g'(\beta \mathbf{z}')))\big] \\
&\leq \frac{\beta}{2}(1-\rho)\min\Big\{\beta, 2\sqrt{2/\pi}\Big\} \\
&\leq \delta^2 \beta \min\Big\{\beta, 2\sqrt{2/\pi}\Big\}.
\end{aligned}$$

Therefore

$$2\kappa + \ln 4 \leq 2n\delta^2\beta \min\{\beta, 2\sqrt{2/\pi}\} + \ln 4 < (d-1)\ln 2. \qquad \blacksquare$$

### 2.2. Low temperatures

The sample complexity lower bound from Theorem 1 tends to 0 as $\beta \to \infty$. This appears to be a limitation of the proof technique, which is only useful for $\beta \lesssim 1/\epsilon$. We next establish a lower bound that improves on Theorem 1 for $\beta \gg 1/\epsilon$.

**Theorem 3** *Fix $\epsilon \in (0,1)$. Assume $\beta \geq 4\sqrt{2/\pi}/\epsilon$. Suppose $\hat{\theta}$ is an estimator that, for any $\theta^\star \in S^{d-1}$,*

$$\Pr(\|\hat{\theta}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) - \theta^\star\| < \epsilon) \geq \frac{1}{2},$$

*where the probability is with respect to the data with distribution determined by $\theta^\star$. Then the sample size $n$ must satisfy*

$$n \geq \frac{d-1}{\epsilon} \cdot \frac{\log(\log(2e/\epsilon)) - \log(16e)}{(1+o(1))\log(2e/\epsilon)}$$

*where the $o(1)$ term depends only $\epsilon$ and vanishes as $\epsilon \to 0$. If $\beta = \infty$, then the sample size must, in fact, satisfy $n \geq (d-1)/(8e\epsilon)$.*

The proof of Theorem 3 essentially follows that of Long (1995) for a lower bound on the sample complexity of PAC learning homogeneous linear classifiers under the uniform distribution on $S^{d-1}$. In Long's setting, the data distribution coincides with ours for $\beta = \infty$. We make a minor modification to the argument to also handle finite (but large) $\beta$. The extra $\log\log(1/\epsilon)/\log(1/\epsilon)$ factor in the finite $\beta$ case appears to be an artifact of the proof technique. We also note that the $\beta = \infty$ case is directly implied by the main result of Long (1995).

Long's proof, as well as ours, relies on the following bound on the shattering number of homogeneous linear classifiers.

**Lemma 4 (Winder, 1966, Corollary on page 816)** *Let $H$ be the family of homogeneous linear classifiers in $\mathbb{R}^d$. For any $x_1, \ldots, x_n \in \mathbb{R}^d$,*

$$|\{(h(x_1), \ldots, h(x_n)) : h \in H\}| \le 2\left(\frac{ne}{d-1}\right)^{d-1}.$$

We also need the following well-known lower bound on the packing number for $S^{d-1}$; the proof is given for completeness.

**Lemma 5** *There exists an $\epsilon$-packing of the unit sphere $S^{d-1}$ with respect to $\ell_2$ distance of cardinality $(1/\epsilon)^{d-1}$.*

**Proof.** Let $W$ be a subspace of $d$-dimensional Euclidean space of dimension $d-1$. By a standard volume argument, there is an $\epsilon$-packing $p_1, \ldots, p_M$ of the unit ball in $W$ with $M \ge (1/\epsilon)^{d-1}$. For each point $p_i$, there is a corresponding point $p_i' \in S^{d-1}$ whose orthogonal projection to $W$ is $p_i$. For any $i \ne j$, we have $\|p_i' - p_j'\| \ge \|p_i - p_j\| \ge \epsilon$, so $p_1', \ldots, p_M'$ is an $\epsilon$-packing of $S^{d-1}$. ■

Finally, we need the following bound on the error rate of the homogeneous linear classifier $h_{\theta^\star}$. (The lemma holds for all $\beta \ge 0$, not just $\beta \gtrsim 1/\epsilon$.)

**Lemma 6** *The error rate of $h_{\theta^\star}$ satisfies*

$$\mathrm{err}_{\theta^\star}(\theta^\star) \le \frac{1}{\beta}\sqrt{\frac{2}{\pi}}.$$

**Proof.** Since

$$\mathrm{err}_{\theta^\star}(\theta^\star) = \Pr(\mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star \le 0) = \mathbb{E}[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)] = \mathbb{E}[g'(-\beta|\mathbf{z}|)]$$

for a normal random variable $\mathbf{z}$, the claim now follows from Lemma 17 (see Appendix A). ■

**Proof of Theorem 3.** Assume without loss of generality that $\epsilon \le 1/2$. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent standard normal random vectors. Let $\mathbf{y}_1^\theta, \ldots, \mathbf{y}_n^\theta$ for all $\theta \in S^{d-1}$ be independent Bernoulli random variables with $\mathbf{y}_i^\theta \mid (\mathbf{x}_1, \ldots, \mathbf{x}_n) \sim \mathrm{Bern}(g'(\beta\mathbf{x}_i^\mathsf{T}\theta))$, and let $\mathbf{Z}^\theta = ((\mathbf{x}_1, \mathbf{y}_1^\theta), \ldots, (\mathbf{x}_n, \mathbf{y}_n^\theta))$. So the data $\mathbf{Z}^\theta$ follows our model with parameter $\theta$, but all $\{\mathbf{Z}^\theta : \theta \in S^{d-1}\}$ share the same $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $\widehat{\mathrm{err}}_\theta(\theta) = (1/n)\sum_{i=1}^n \mathbb{1}\{h_\theta(\mathbf{x}_i) \ne \mathbf{y}_i^\theta\}$ be the empirical error rate of $h_\theta$ (with respect to $\mathbf{Z}^\theta$).

Suppose $\Pr(\|\hat\theta(\mathbf{Z}^\theta) - \theta\| < \epsilon) \ge 1/2$ for all $\theta \in S^{d-1}$. By Lemma 6, since $\beta \ge 4\sqrt{2/\pi}/\epsilon$, we have

$$\mathrm{err}_\theta(\theta) \le \frac{\epsilon}{4}.$$

Let $U$ be a $2\epsilon$-packing of $S^{d-1}$ with respect to $\ell_2$ of cardinality $|U| \ge (2\epsilon)^{-(d-1)}$, as guaranteed to exist by Lemma 5. Let $G_\theta$ be the (indicator of) the event $\|\hat\theta(\mathbf{Z}^\theta) - \theta\| < \epsilon$ and $\widehat{\mathrm{err}}_\theta(\theta) \le \epsilon$. By Markov's inequality,

$$\Pr(\widehat{\mathrm{err}}_\theta(\theta) \ge \epsilon) \le \frac{1}{4}.$$

So

$$\mathbb{E}\left[\sum_{\theta\in U} G_\theta\right] = \sum_{\theta\in U}\mathbb{E}[G_\theta] \ge \sum_{\theta\in U}\left(1 - \frac{1}{2} - \frac{1}{4}\right) = \frac{|U|}{4} \ge \frac{1}{4}\left(\frac{1}{2\epsilon}\right)^{d-1}.$$

Consider any two $\theta, \theta' \in U$. If $\mathbf{Z}^\theta = \mathbf{Z}^{\theta'}$, then either $\|\hat{\theta}(\mathbf{Z}^\theta) - \theta\| > \epsilon$ or $\|\hat{\theta}(\mathbf{Z}^{\theta'}) - \theta'\| > \epsilon$ (since $\|\theta - \theta'\| \geq 2\epsilon$), so at least one of $G_\theta$ and $G_{\theta'}$ is zero. Moreover, if $G_\theta = 1$, then the labels $\mathbf{y}_i^\theta$ are realized by the homogeneous linear classifier determined by $\theta$—except for up to $\lfloor n\epsilon \rfloor$ labels, which could be flipped. Therefore $\mathbb{E}\left[\sum_{\theta \in U} G_\theta\right]$ is at most the number of ways of labeling $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ by homogeneous linear classifiers determined by weight vectors from $U$, multiplied by $2^{\lfloor n\epsilon \rfloor}\binom{n}{\lfloor n\epsilon \rfloor}$. Hence, by Lemma 4,

$$\mathbb{E}\left[\sum_{\theta \in U} G_\theta\right] \leq 2\left(\frac{ne}{d-1}\right)^{d-1} \cdot \left(\frac{2ne}{n\epsilon}\right)^{n\epsilon} = 2\left(\frac{e}{M\epsilon} \cdot \left(\frac{2e}{\epsilon}\right)^{1/M}\right)^{d-1}$$

where $M = (d-1)/(n\epsilon)$. Combining the upper and lower bounds on $\mathbb{E}\left[\sum_{\theta \in U} G_\theta\right]$ gives

$$\frac{e}{M\epsilon} \cdot \left(\frac{2e}{\epsilon}\right)^{1/M} \geq \frac{1}{8^{1/(d-1)}} \cdot \frac{1}{2\epsilon} \geq \frac{1}{16\epsilon}.$$

Taking logarithms of both sides and simplifying gives the following inequality:

$$\left(\frac{M}{16e}\right) \log\left(\frac{M}{16e}\right) \leq \frac{1}{16e} \log \frac{2e}{\epsilon}.$$

Let $T(\epsilon) = \log(2e/\epsilon)/(16e)$. Then, using asymptotic expansion of the product log function (Corless et al., 1996), we have

$$\frac{M}{16e} \leq \frac{T(\epsilon)}{\log T(\epsilon)}(1 + o(1)),$$

where the $o(1)$ term vanishes as $T(\epsilon) \to \infty$ (i.e., $\epsilon \to 0$). This implies

$$n \geq \frac{d-1}{16e\epsilon} \cdot \frac{\log T(\epsilon)}{(1 + o(1))T(\epsilon)}$$

as claimed.

If $\beta = \infty$, then $\Pr(\widehat{\mathrm{err}}_\theta(\theta) = 0) = 1$, and hence we have

$$\frac{1}{2}\left(\frac{1}{2\epsilon}\right)^{d-1} \leq \mathbb{E}\left[\sum_{\theta \in U} G_\theta\right] \leq 2\left(\frac{ne}{d-1}\right)^{d-1}.$$

Therefore

$$n \geq \frac{d-1}{8e\epsilon}.$$

■

## 3. Upper bounds on the sample complexity

In this section, we give upper bounds on the sample complexity based on three different estimators for the three different regimes of $\beta$. Throughout this section, we fix a "ground truth" parameter $\theta^\star \in S^{d-1}$ determining the distribution of $(\mathbf{x}, \mathbf{y})$.

### 3.1. High temperatures

The analysis of Plan et al. (2017, Corollary 3.5) implies, for any $\epsilon \in (0,1)$, the "linear estimator" $\hat{\theta}_{\mathrm{linear}}$ of Plan and Vershynin (2012) (or the "Average" algorithm of Servedio (1999))

$$\hat{\theta}_{\mathrm{linear}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) = \operatorname*{arg\,max}_{\theta \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{x}_i^\mathsf{T} \theta$$

satisfies $\|\hat{\theta}_{\mathrm{linear}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) - \theta^\star\| \le \epsilon$ in expectation provided that

$$n \ge C \max\left\{\frac{1}{\beta^2}, 1\right\} \frac{d}{\epsilon^2}, \tag{1}$$

where $C > 0$ is a universal positive constant. This matches the lower bound from Theorem 1 (up to constants) when $\beta \lesssim 1$.

### 3.2. Moderate and low temperatures

When $\beta \gtrsim 1$, the sample size requirement of the linear estimator in (1) has a suboptimal dependence on $\beta$. In particular, the sample size requirement does not become smaller as $\beta$ becomes larger.

In this section, we analyze a different estimator, the empirical ReLU risk minimizer $\hat{\theta}_{\mathrm{relu}}$:

$$\hat{\theta}_{\mathrm{relu}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) \in \operatorname*{arg\,min}_{\theta \in S^{d-1}} \frac{1}{n} \sum_{i=1}^n [-\mathbf{y}_i \mathbf{x}_i^\mathsf{T} \theta]_+, \tag{2}$$

where $[x]_+ = \max\{0, x\}$ (the rectified linear function used in ReLUs). Notice that if $[\cdot]_+$ was omitted in each summand of (2), then the estimator $\hat{\theta}_{\mathrm{relu}}$ would be the same as $\hat{\theta}_{\mathrm{linear}}$. This estimator is also related to the Perceptron algorithm (Rosenblatt, 1958), since the latter can be viewed as a stochastic subgradient method for minimizing the empirical ReLU risk. However, we do not know if such subgradient methods minimize the objective over the nonconvex domain $S^{d-1}$.

**Theorem 7** *Assume $\beta \ge 1 + c$ for some positive constant $c > 0$. Fix any $\epsilon, \delta \in (0,1)$, and suppose*

$$n \ge C\left(\frac{d\log(d/\epsilon) + \log(1/\delta)}{\beta\epsilon^2} + \frac{d\log(d/\epsilon) + \log(1/\delta)}{\epsilon}\right)$$

*where $C > 0$ is an absolute constant. Then with probability at least $1 - \delta$, we have*

$$\|\hat{\theta}_{\mathrm{relu}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) - \theta^\star\| \le \epsilon.$$

The main technical work in the proof of Theorem 7 is understanding the concentration properties of random variables of the form

$$\boldsymbol{\delta}_\theta := [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta]_+ - [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star]_+, \quad \theta \in S^{d-1}.$$

This was studied by Kuchelmeister and van de Geer (2023) in the case of the probit regression model. We obtain the necessary moment bounds for the logistic regression model. The following lemma (proved in Appendix B) gives the required bounds.

**Lemma 8** *For any $\theta \in S^{d-1}$, and any integer $q \geq 2$,*

$$\mathbb{E}[|\boldsymbol{\delta}_\theta|^q] \leq \frac{q!v}{2} b^{q-2}$$

*where*

$$b = C\|\theta - \theta^\star\| \quad \textit{and} \quad v \leq C\|\theta - \theta^\star\|^2 \left(\frac{1}{\beta} + \|\theta - \theta^\star\|\right)$$

*and $C > 0$ is an absolute constant.*

Given Lemma 8, the rest of the analysis is mostly standard. Lemma 8 and Bernstein's inequality implies that each random variable in the empirical process above is concentrated around its expectation.

**Lemma 9** *For any $\theta \in S^{d-1}$, define $\boldsymbol{\delta}_\theta^i := [-\mathbf{y}_i \mathbf{x}_i^\mathsf{T}\theta]_+ - [-\mathbf{y}_i \mathbf{x}_i^\mathsf{T}\theta^\star]_+$ for all $i = 1, \ldots, n$. For any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have*

$$\mathbb{E}[\boldsymbol{\delta}_\theta] - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_\theta^i \leq C\|\theta - \theta^\star\| \sqrt{\left(\frac{1}{\beta} + \|\theta - \theta^\star\|\right)\frac{\log(1/\delta)}{n}} + C\|\theta - \theta^\star\|\frac{\log(1/\delta)}{n},$$

*where $C > 0$ is an absolute constant.*

To bound all random variables in the stochastic process simultaneously, we use a covering argument, again following Kuchelmeister and van de Geer (2023).

**Lemma 10 (Kuchelmeister and van de Geer, 2023, Lemma 5.1.2)** *For any $\varepsilon_0 \in (0,1)$ and any $A \subseteq S^{d-1}$, let $T(A, \varepsilon_0)$ be an $\varepsilon_0$-cover of $A$ with respect to $\ell_2$ distance. Define $\boldsymbol{\delta}_\theta^i := [-\mathbf{y}_i \mathbf{x}_i^\mathsf{T}\theta]_+ - [-\mathbf{y}_i \mathbf{x}_i^\mathsf{T}\theta^\star]_+$ for all $i = 1, \ldots, n$ and all $\theta \in A$. For any $\delta \in (0,1)$, with probability at least $1 - \delta$, we have*

$$\sup_{\theta \in A}\left\{\mathbb{E}[\boldsymbol{\delta}_\theta] - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_\theta^i\right\} \leq \max_{\theta \in T(A,\varepsilon_0)}\left\{\mathbb{E}[\boldsymbol{\delta}_\theta] - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_\theta^i\right\} + \varepsilon_0\left(\sqrt{\frac{2\ln(1/\delta)}{n}} + 2\sqrt{d}\right).$$

To relate $\mathbb{E}[\boldsymbol{\delta}_\theta]$ to $\|\theta - \theta^\star\|$, we use the following lemma.

**Lemma 11** *For any $\theta \in S^{d-1}$, we have*

$$\mathbb{E}[\boldsymbol{\delta}_\theta] \geq \frac{1}{8}\sqrt{\frac{2}{\pi}}\left(1 - \frac{1}{\beta^2}\right)\|\theta - \theta^\star\|^2.$$

**Proof.** The expected KL divergence of $\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta))$ from $\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta^\star))$ is the expected excess logistic loss of $w := \beta\theta$ compared to $w^\star := \beta\theta^\star$:

$$\mathbb{E}\Big[\mathrm{KL}(\mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta)) \| \mathrm{Bern}(g'(\beta \mathbf{x}^\mathsf{T}\theta^\star)))\Big] = \mathbb{E}\Big[\ln(1 + e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w})\Big] - \mathbb{E}\Big[\ln(1 + e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w^\star})\Big]$$

(recalling that the conditional distribution of $\mathbf{y}$ given $\mathbf{x} = x$ is $\mathrm{Bern}(g'(x^\mathsf{T}w^\star))$). Furthermore, the logistic loss $w \mapsto \ln(1 + e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w})$ decomposes into the sum of a label-dependent part and a label-independent part:

$$\ln(1 + e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w}) = [-\mathbf{y}\mathbf{x}^\mathsf{T}w]_+ + \ln(1 + e^{-|\mathbf{x}^\mathsf{T}w|}).$$

(This decomposition is the same as that from Kuchelmeister and van de Geer (2023), and a similar decomposition was used by Bach (2010).) We can thus write the excess expected logistic loss as

$$\begin{aligned}\mathbb{E}\Big[\ln(1+e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w})\Big]-\mathbb{E}\Big[\ln(1+e^{-\mathbf{y}\mathbf{x}^\mathsf{T}w^\star})\Big] &= \mathbb{E}\big[[-\mathbf{y}\mathbf{x}^\mathsf{T}w]_+\big]+\mathbb{E}\Big[\ln(1+e^{-|\mathbf{x}^\mathsf{T}w|})\Big]\\ &\quad-\mathbb{E}\big[[-\mathbf{y}\mathbf{x}^\mathsf{T}w^\star]_+\big]-\mathbb{E}\Big[\ln(1+e^{-|\mathbf{x}^\mathsf{T}w^\star|})\Big]\\ &= \mathbb{E}\big[[-\mathbf{y}\mathbf{x}^\mathsf{T}w]_+-[-\mathbf{y}\mathbf{x}^\mathsf{T}w^\star]_+\big]\end{aligned}$$

since $\mathbf{x}^\mathsf{T}w$ and $\mathbf{x}^\mathsf{T}w^\star$ have the same distribution. Therefore, we have

$$\mathbb{E}\big[[-\mathbf{y}\mathbf{x}^\mathsf{T}\theta]_+-[-\mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star]_+\big]=\frac{1}{\beta}\,\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta\mathbf{x}^\mathsf{T}\theta))\|\,\mathrm{Bern}(g'(\beta\mathbf{x}^\mathsf{T}\theta^\star)))\big].$$

The claim now follows by applying Lemma 19 (see Appendix A). ■

**Proof of Theorem 7.** Fix some $\varepsilon_0, r_0 \in (0,1)$ and let $r_j = 2^j r_0$ for $1 \le j \le J := \lceil \log_2(2/r_0)\rceil$. Define $A_0 = \{\theta \in S^{d-1} : \|\theta-\theta^\star\| \le r_0\}$ and $A_j = \{\theta \in S^{d-1} : r_{j-1} < \|\theta-\theta^\star\| \le r_j\}$ for all $j \ge 1$. Let $T_j = T(A_j, \varepsilon_0(r_j/2)^{3/2})$ be an $\varepsilon_0(r_j/2)^{3/2}$-cover of $A_j$ with respect to $\ell_2$ distance of cardinality at most $(2(2/r_j)^{3/2}\varepsilon_0^{-1})^{d-1}$ (cf. Lemma 5). Fix $\delta_0 \in (0, 1/(J+\sum_{j=0}^J|T_j|))$, and apply Lemma 10 to each $A_j$ for $1 \le j \le J$, and Lemma 9 to each $\theta \in T$. By a union bound, with probability at least $1-(J+\sum_{j=0}^J|T_j|)\delta_0$, we have for all $\theta \in S^{d-1}$,

$$\begin{aligned}\mathbb{E}[\boldsymbol{\delta}_\theta] \le \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_\theta^i + 2C'\Delta_{r_0}(\theta)\sqrt{\left(\frac{1}{\beta}+\Delta_{r_0}(\theta)\right)\frac{\log(1/\delta_0)}{n}} + 2C'\Delta_{r_0}(\theta)\frac{\log(1/\delta_0)}{n}\\ + \varepsilon_0\Delta_{r_0}(\theta)^{3/2}\left(\sqrt{\frac{2\ln(1/\delta_0)}{n}}+2\sqrt{d}\right)\end{aligned}$$

where $\Delta_{r_0}(\theta) := \max\{r_0, \|\theta-\theta^\star\|\}$, $C'>0$ is some absolute constant, and $\boldsymbol{\delta}_\theta^i$ are as in Lemma 9. We condition on the event that these inequalities hold for all $\theta \in S^{d-1}$. Since $\hat\theta_{\mathrm{relu}}$ minimizes $\frac{1}{n}\sum_{i=1}^n[-\mathbf{y}_i\mathbf{x}_i^\mathsf{T}\theta]_+$, it follows that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}^i_{\hat\theta_{\mathrm{relu}}} \le 0$. Therefore, together with Lemma 11 to lower-bound $\mathbb{E}[\boldsymbol{\delta}_\theta]$ for $\theta=\hat\theta_{\mathrm{relu}}$, we have

$$\begin{aligned}\frac{1}{8}\sqrt{\frac{2}{\pi}}\left(1-\frac{1}{\beta^2}\right)\|\hat\theta_{\mathrm{relu}}-\theta^\star\|^2 \le 2C'\Delta_{r_0}(\hat\theta_{\mathrm{relu}})\sqrt{\frac{\log(1/\delta_0)}{\beta n}} + 2C'\Delta_{r_0}(\hat\theta_{\mathrm{relu}})^{3/2}\sqrt{\frac{\log(1/\delta_0)}{n}}\\ + 2C'\Delta_{r_0}(\hat\theta_{\mathrm{relu}})\frac{\log(1/\delta_0)}{n} + \varepsilon_0\Delta_{r_0}(\hat\theta_{\mathrm{relu}})^{3/2}\left(\sqrt{\frac{2\ln(1/\delta_0)}{n}}+2\sqrt{d}\right).\end{aligned}$$

Let $r_0=\epsilon$, and suppose that $\|\hat\theta_{\mathrm{relu}}-\theta^\star\| > \epsilon$, in which case we have $\Delta_{r_0}(\hat\theta_{\mathrm{relu}}) = \|\hat\theta_{\mathrm{relu}}-\theta^\star\|$. Then, the display above has $\|\hat\theta_{\mathrm{relu}}-\theta^\star\|$ on both sides of the inequality, and it simplifies to an inequality in $x = \|\hat\theta_{\mathrm{relu}}-\theta^\star\|^{1/2}$ of the form $x - b\sqrt{x} - c \le 0$ for some $b, c \ge 0$, which in turn implies $x \le 1.5(b^2+c)$. Therefore

$$\|\hat\theta_{\mathrm{relu}}-\theta^\star\| \le C''\left(\varepsilon_0^2\left(\frac{\log(1/\delta_0)}{n}+d\right)+\sqrt{\frac{\log(1/\delta_0)}{\beta n}}+\frac{\log(1/\delta_0)}{n}\right)$$

for some absolute constant $C'' > 0$. (Recall that we have assumed $\beta \geq 1 + c$ for some absolute constant $c > 0$, and hence $1 - 1/\beta^2 \geq c(2+c)/(1+c)^2$.) Plug-in $\varepsilon_0 = \sqrt{\epsilon/(4C''d)}$, $\delta_0 = \delta/(J + \sum_{j=0}^{J} |T_j|)$, and

$$n \geq \frac{16 \log(1/\delta_0)}{\beta \epsilon^2} + \frac{4C'' \log(1/\delta_0)}{\epsilon}$$

to conclude that $\|\hat{\theta}_{\mathrm{relu}} - \theta^\star\| \leq \epsilon$. The lower bound $n$ comes from the assumption in the theorem statement, and the fact that $J = O(\log(1/\epsilon))$ and $\sum_{j=0}^{J} |T_j| = O(d/\epsilon)^{O(d)}$. ■

### 3.3. Low temperatures

The objective function minimized by $\hat{\theta}_{\mathrm{relu}}$ in (2) uses the magnitude of the inner product $\mathbf{x}_i^{\mathsf{T}}\theta$ whenever its sign differs from that of $\mathbf{y}_i$. At sufficiently low temperatures ($\beta \gtrsim 1/\epsilon$), this magnitude information can be safely ignored. Specifically, we show that the empirical (zero-one loss) risk minimizer

$$\hat{\theta}_{\mathrm{ERM}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) \in \operatorname*{arg\,min}_{\theta \in S^{d-1}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\mathbf{y}_i \mathbf{x}_i^{\mathsf{T}} \theta \leq 0\}$$

achieves near-optimal sample complexity in this regime.

**Theorem 12** *Fix any $\epsilon, \delta \in (0,1)$, and assume*

$$\beta \geq \frac{4\sqrt{2\pi}}{\epsilon}$$

*and*

$$n \geq \frac{C(d \log(1/\epsilon) + \log(1/\delta))}{\epsilon}$$

*where $C > 0$ is an absolute constant. Then with probability at least $1 - \delta$, we have*

$$\|\hat{\theta}_{\mathrm{ERM}}((\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n) - \theta^\star\| \leq \epsilon.$$

Note that the sample size requirement given in Theorem 12 removes a $\log(d)$ factor from that in Theorem 7 in the low temperature regime.

The proof of Theorem 12 is largely based on the following standard performance guarantee for empirical risk minimization, combined with the fact that the VC dimension of the class of homogeneous linear classifiers is $d$.[2]

**Lemma 13 (Vapnik and Chervonenkis, 1971)** *There is a universal constant $C > 0$ such that the following holds. Let $\mathrm{opt} = \min_{\theta \in S^{d-1}} \mathrm{err}_{\theta^\star}(\theta)$. For any $\varepsilon \in (0,1)$ and $\delta \in (0,1)$, if*

$$n \geq C\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2} \mathrm{opt} + \frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right),$$

2. Note that Lemma 13 is true under every distribution for $(\mathbf{x}, \mathbf{y})$; it does not rely on specific properties of our model. In the special case of where $\mathbf{x}$ is the $d$-dimensional standard normal (or any other spherically symmetric distribution) and $\beta = \infty$ (which implies $\mathrm{opt} = 0$), the $\log(1/\varepsilon)$ can be removed (Long, 1995). So, in this case, the sample size requirement is $C(d + \log(1/\delta))/\varepsilon$.

*then the empirical risk minimizer* $\hat{\theta}_{\mathrm{ERM}}$ *satisfies*

$$\mathrm{err}_{\theta^\star}(\hat{\theta}_{\mathrm{ERM}}) - \mathrm{opt} \le \varepsilon$$

*with probability at least* $1 - \delta$ *over the realization of the data.*

The following lemma (proved in Appendix C) relates $\|\hat{\theta}_{\mathrm{ERM}} - \theta^\star\|$ to $\mathrm{err}_{\theta^\star}(\hat{\theta}_{\mathrm{ERM}}) - \mathrm{err}_{\theta^\star}(\theta^\star)$.

**Lemma 14** *For any* $\theta \in S^{d-1}$,

$$\|\theta - \theta^\star\| \le \pi(\mathrm{err}_{\theta^\star}(\theta) - \mathrm{err}_{\theta^\star}(\theta^\star)) + \frac{2\sqrt{2\pi}}{\beta}.$$

**Proof of Theorem 12.** Note that

$$\min_{\theta \in S^{d-1}} \mathrm{err}_{\theta^\star}(\theta) = \mathrm{err}_{\theta^\star}(\theta^\star).$$

Therefore, we can combine Lemma 13 (with $\varepsilon = \epsilon/(2\pi)$) and Lemma 14 with the bound on $\mathrm{opt} = \mathrm{err}_{\theta^\star}(\theta^\star)$ from Lemma 6 to obtain the desired bound on $\|\hat{\theta}_{\mathrm{ERM}} - \theta^\star\|$. ■

### 3.4. Adaptivity

If the inverse temperature $\beta$ is unknown, then we need to determine which of the aforementioned estimators to use in a data-driven fashion. Notice that the estimators $\hat{\theta}_{\mathrm{linear}}$, $\hat{\theta}_{\mathrm{relu}}$, and $\hat{\theta}_{\mathrm{ERM}}$ may be computed without explicit knowledge of $\beta$. Therefore, it suffices to (coarsely) distinguish between the high ($\beta \lesssim 1$) and moderate-or-low ($\beta \gtrsim 1$) temperature regimes. This can be done using an estimate of $\mathrm{err}_{\theta^\star}(\theta^\star)$ (e.g., training error rate of $\hat{\theta}_{\mathrm{ERM}}$) and reasoning about its relationship to $\beta$.

## 4. Discussion

Our characterization of the sample complexity of estimation in logistic regression delineates the high, moderate, and low temperature regimes. However, we are only aware of computationally efficient estimators that achieve the (near) optimal sample complexities at high and zero temperatures: e.g., the linear estimator of Plan and Vershynin (2012) for $\beta \lesssim 1$, and the estimator based on solving a linear feasibility program (or the algorithm of Balcan and Long (2013)) for $\beta = \infty$. Note that, although the ReLU loss is convex, we need to minimize it over the sphere. It would be interesting to determine if the MLE itself (i.e., minimizing the logistic loss), or its efficient approximations, can shown to achieve optimal sample complexity.

## Acknowledgements

We are most grateful to Akshay Krishnamurthy for helpful discussions about non-asymptotic analysis of MLE, and to the anonymous COLT reviewers for their constructive feedback and suggestions. Part of this work was done while DH was visiting the Halıcıoğlu Data Science Institute at UC San Diego in Spring 2023. We acknowledge the support of the National Science Foundation under grants IIS 2040971 and CCF 2217058/2133484.

## References

Francis Bach. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4: 384–414, 2010.

Maria-Florina Balcan and Phil Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316, 2013.

Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *Conference on Information Sciences and Systems*, pages 16–21, 2008.

Xi Chen, Adityanand Guntuboyina, and Yuchen Zhang. On Bayes risk lower bounds. *Journal of Machine Learning Research*, 17(1):7687–7744, 2016.

Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. On the Lambert W function. *Advances in Computational Mathematics*, 5:329–359, 1996.

Luc Devroye and Gábor Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7):1011–1018, 1995.

William Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. John Wiley & Sons, 3rd edition, 1968.

Dylan J Foster, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Logistic regression: The importance of being improper. In *Conference on Learning Theory*, pages 167–208, 2018.

Te Sun Han and Sergio Verdú. Generalizing the Fano inequality. *IEEE Transactions on Information Theory*, 40(4):1247–1251, 1994.

Elad Hazan, Tomer Koren, and Kfir Y Levy. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory*, pages 197–209, 2014.

Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *Journal of Multivariate Analysis*, 73(1):120–135, 2000.

Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory*, 59(4):2082–2102, 2013.

Sham M Kakade and Andrew Ng. Online bounds for Bayesian algorithms. In *Advances in Neural Information Processing Systems 17*, 2004.

Felix Kuchelmeister and Sara van de Geer. Finite sample rates for logistic regression with small noise or few samples. *arXiv preprint arXiv:2305.15991*, 2023.

Philip M Long. On the sample complexity of PAC learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

Philip M Long. An upper bound on the sample complexity of PAC-learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

Enno Mammen and Alexandre B Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Jaouad Mourtada and Stéphane Gaïffas. An improper estimator with optimal excess risk in misspecified density estimation and logistic regression. *Journal of Machine Learning Research*, 23 (31):1–49, 2022.

Dmitrii M Ostrovskii and Francis Bach. Finite-sample analysis of M-estimators using self-concordance. *Electronic Journal of Statistics*, 15:326–391, 2021.

Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.

Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *Information and Inference: A Journal of the IMA*, 6(1):1–40, 2017.

Stephen Portnoy. Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *The Annals of Statistics*, pages 356–366, 1988.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems 32*, 2019.

Rocco A Servedio. On PAC learning using Winnow, Perceptron, and a Perceptron-like algorithm. In *Conference on Computational Learning Theory*, pages 296–307, 1999.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems 23*, 2010.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Vladimir Naumovich Vapnik and Alexey Yakovlevich Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.

Robert O Winder. Partitions of $N$-space by hyperplanes. *SIAM Journal on Applied Mathematics*, 14(4):811–818, 1966.

Tong Zhang. Information-theoretic upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 52(4):1307–1321, 2006.

## Appendix A. Bounds on Gaussian integrals

**Lemma 15 (Feller, 1968, page 175)** *Let* $\mathbf{z} \sim \mathrm{N}(0,1)$. *For any* $t > 0$,

$$\left(1 - \frac{1}{t^2}\right)\frac{1}{t\sqrt{2\pi}}\exp(-t^2/2) \leq \Pr(\mathbf{z} \geq t) \leq \frac{1}{t\sqrt{2\pi}}\exp(-t^2/2).$$

Recall that $g(\eta) = \ln(1+\exp(\eta))$.

**Lemma 16** *Let* $\mathbf{z} \sim \mathrm{N}(0,1)$. *For any* $\beta > 0$,

$$\frac{1}{4}\,\mathbb{E}[\exp(-\beta|\mathbf{z}|)] \leq \mathbb{E}[g''(\beta\mathbf{z})] \leq \min\left\{\frac{1}{2},\ \mathbb{E}[\exp(-\beta|\mathbf{z}|)]\right\} \tag{3}$$

*and*

$$\frac{1}{\beta}\left(1 - \frac{1}{\beta^2}\right)\sqrt{\frac{2}{\pi}} \leq \mathbb{E}[\exp(-\beta|\mathbf{z}|)] \leq \frac{1}{\beta}\sqrt{\frac{2}{\pi}}. \tag{4}$$

**Proof.** First, observe that

$$g''(\eta) = \frac{1}{(1+\exp(\eta))(1+\exp(-\eta))} = \frac{1}{(1+\exp(|\eta|))(1+\exp(-|\eta|))} \leq \min\left\{\frac{1}{2},\ \exp(-|\eta|)\right\},$$

while

$$\frac{\exp(|\eta|)}{(1+\exp(\eta))(1+\exp(-\eta))} = \left(\frac{1}{1+\exp(-|\eta|)}\right)^2 \geq \frac{1}{4}.$$

Plugging in $\eta = \beta\mathbf{z}$ and taking expectations gives the upper- and lower-bounds on $\mathbb{E}[g''(\beta\mathbf{z})]$. For the upper- and lower-bounds on $\mathbb{E}[\exp(-\beta|\mathbf{z}|)]$, we have

$$\begin{aligned}
\mathbb{E}[\exp(-\beta|\mathbf{z}|)] &= 2\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp(-\beta z - z^2/2)\,\mathrm{d}z \\
&= 2\exp(\beta^2/2)\int_0^\infty \frac{1}{\sqrt{2\pi}}\exp(-(z+\beta)^2/2)\,\mathrm{d}z \\
&= 2\exp(\beta^2/2)\Pr(\mathbf{z} \geq \beta),
\end{aligned}$$

and therefore the conclusion follows by applying Lemma 15. ■

**Lemma 17** *Let* $\mathbf{z} \sim \mathrm{N}(0,1)$. *For any* $\beta > 0$,

$$\mathbb{E}[g'(-\beta|\mathbf{z}|)] \leq \min\left\{\frac{1}{2},\ \frac{1}{2} - \frac{\beta}{4}\sqrt{\frac{2}{\pi}}\left(1 - \frac{\beta^2}{6}\right),\ \frac{1}{\beta}\sqrt{\frac{2}{\pi}}\right\}.$$

**Proof.** First, observe that $g'(\eta) \leq 1/2$ for all $\eta \leq 0$. Furthermore, by Taylor's theorem, for any $\eta \in \mathbb{R}$, there exists $h \in \mathbb{R}$ (between 0 and $\eta$) such that

$$\begin{aligned}
g'(\eta) &= g'(0) + g''(\eta)\eta + \frac{1}{2}g'''(\eta)\eta^2 + \frac{1}{6}g''''(h)\eta^3 \\
&\leq g'(0) + g''(0)\eta + \frac{1}{2}g'''(0)\eta^2 + \frac{1}{48}|\eta|^3 \\
&= \frac{1}{2} + \frac{1}{4}\eta + \frac{1}{48}|\eta|^3,
\end{aligned}$$

where the inequality uses the fact that $|g''''(h)| \le 1/8$ for all $h \in \mathbb{R}$. Therefore

$$\begin{aligned}\mathbb{E}[g'(-\beta|\mathbf{z}|)] &\le \frac{1}{2} - \frac{\beta}{4}\,\mathbb{E}[|\mathbf{z}|] + \frac{\beta^3}{48}\,\mathbb{E}[|\mathbf{z}|^3] \\ &= \frac{1}{2} - \frac{\beta}{4}\sqrt{\frac{2}{\pi}} + \frac{\beta^3}{24}\sqrt{\frac{2}{\pi}} \\ &= \frac{1}{2} - \frac{\beta}{4}\sqrt{\frac{2}{\pi}}\left(1 - \frac{\beta^2}{6}\right).\end{aligned}$$

Finally, we also have

$$\mathbb{E}[g'(-\beta|\mathbf{z}|)] = \mathbb{E}\left[\frac{1}{1+\exp(\beta|\mathbf{z}|)}\right] \le \mathbb{E}[\exp(-\beta|\mathbf{z}|)] \le \frac{1}{\beta}\sqrt{\frac{2}{\pi}},$$

where the final inequality follows by Lemma 16. ■

**Lemma 18** *Let* $\mathbf{z} \sim \mathrm{N}(0,1)$. *For any* $\beta > 0$ *and any non-negative integer* $q$,

$$\mathbb{E}[\exp(-\beta|\mathbf{z}|)|\mathbf{z}|^q] \le \frac{q!}{\beta^q}.$$

We remark that the bound in Lemma 18 can be improved to $O(q!/\beta^{q+1})$, which is tight up to constants. This is because $\mathbb{E}[\exp(-\beta|\mathbf{z}|)|\mathbf{z}|^q] = \sqrt{2/\pi}\mu_q m_\beta$, where $\mu_q$ is the $q$-th uncentered moment of the $[0,\infty)$-truncated $\mathrm{N}(-\beta,1)$ distribution, and $m_\beta = (1-\Phi(\beta))/\phi(\beta)$ is the Mills ratio for $\mathrm{N}(0,1)$ at $\beta$. However, we do not need this improved bound in the present work.

**Proof of Lemma 18.** The function $f\colon \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x) = -\beta x + q\log(x)$ is maximized at $x = q/\beta$. Therefore

$$\begin{aligned}\mathbb{E}[\exp(-\beta|\mathbf{z}|)|\mathbf{z}|^q] &= \mathbb{E}[\exp(-\beta|\mathbf{z}| + q\ln|\mathbf{z}|)] \\ &\le \exp(-\beta(q/\beta) + q\ln(q/\beta)) \\ &= \exp(-q + q\ln(q) - q\ln(\beta)) \\ &\le \exp(\ln(q!) - q\ln(\beta)) \\ &= \frac{q!}{\beta^q}.\end{aligned}$$

■

Recall that the Bernoulli distributions form an exponential family $\{p_\eta : \eta \in \mathbb{R}\}$, where

$$p_\eta(y) = \exp(\eta \mathbb{1}\{y=1\} - g(\eta))$$

and $g(\eta) = \ln(1+\exp(\eta))$ is the log partition function for $p_\eta$. The mean parameter—i.e., the mean of $\mathbb{1}\{y=1\}$ under $p_\eta$—is given by $g'(\eta)$. In the proof of the following lemma, we use the fact that the KL divergence $\mathrm{KL}(p_\eta \| p_{\eta'})$ can be expressed as a Bregman divergence associated with $g$:

$$\mathrm{KL}(p_\eta \| p_{\eta'}) = g(\eta') - \big(g(\eta) + g'(\eta)(\eta' - \eta)\big).$$

**Lemma 19** *Let* $\mathbf{z}$ *and* $\mathbf{z}'$ *be* $\mathrm{N}(0,1)$ *random variables with correlation* $\rho$*. Then*

$$\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta\mathbf{z}))\|\,\mathrm{Bern}(g'(\beta\mathbf{z}')))\big] \le \frac{\beta}{2}(1-\rho)\min\Big\{\beta, 2\sqrt{2/\pi}\Big\}$$

*and*

$$\mathbb{E}\big[\mathrm{KL}(\mathrm{Bern}(g'(\beta\mathbf{z}))\|\,\mathrm{Bern}(g'(\beta\mathbf{z}')))\big] \ge \frac{\beta}{4}(1-\rho)\left(1-\frac{1}{\beta^2}\right)\sqrt{\frac{2}{\pi}}.$$

**Proof.** For any $\eta, \eta' \in \mathbb{R}$, we have

$$\mathrm{KL}(\mathrm{Bern}(g'(\eta))\|\,\mathrm{Bern}(g'(\eta'))) = g(\eta') - g(\eta) - g'(\eta)(\eta'-\eta).$$

Hence,

$$\begin{aligned}\mathbb{E}[\mathrm{KL}(\mathrm{Bern}(g'(\beta\mathbf{z}))\|\,\mathrm{Bern}(g'(\beta\mathbf{z}')))] &= \mathbb{E}[g(\beta\mathbf{z}') - g(\beta\mathbf{z}) - g'(\beta\mathbf{z})\beta(\mathbf{z}'-\mathbf{z})] \\ &= \mathbb{E}[g'(\beta\mathbf{z})\beta(\mathbf{z}-\mathbf{z}')].\end{aligned}$$

Since $\mathbf{z}$ and $\mathbf{z}'$ have correlation $\rho$, we may write

$$\mathbf{z}' = \rho\mathbf{z} + \sqrt{1-\rho^2}\mathbf{z}_\perp,$$

where $\mathbf{z}_\perp \sim \mathrm{N}(0,1)$ is independent of $\mathbf{z}$. Thus

$$\begin{aligned}\mathbb{E}[g'(\beta\mathbf{z})\beta(\mathbf{z}-\mathbf{z}')] &= \mathbb{E}\Big[g'(\beta\mathbf{z})\beta\Big((1-\rho)\mathbf{z} - \sqrt{1-\rho^2}\mathbf{z}_\perp\Big)\Big] \\ &= \beta(1-\rho)\,\mathbb{E}[g'(\beta\mathbf{z})\mathbf{z}] \\ &= \beta^2(1-\rho)\,\mathbb{E}[g''(\beta\mathbf{z})],\end{aligned}$$

where the final step follows by Stein's identity. Now apply Lemma 16 to obtain the conclusion. ■

## Appendix B. Proof of Lemma 8

The following lemma is implicit in the proofs of Lemma 5.2.1 and Lemma A.2.1 of Kuchelmeister and van de Geer (2023).

**Lemma 20 (Kuchelmeister and van de Geer, 2023)** *Fix* $\theta^\star, \theta \in S^{d-1}$ *and* $p\colon \mathbb{R} \to [0,1]$ *satisfying* $p(-t) = 1 - p(t)$ *for all* $t \in \mathbb{R}$*. Let* $\mathbf{x}$ *be a standard normal random vector in* $\mathbb{R}^d$*; let the conditional distribution of* $\mathbf{y}$ *given* $\mathbf{x}$ *be* $\mathrm{Bern}(p(\mathbf{x}^\mathsf{T}\theta^\star))$*; and define*

$$\boldsymbol{\delta}_\theta := [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta]_+ - [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star]_+.$$

*For any integer* $q \ge 2$*,*

$$\begin{aligned}\mathbb{E}[|\boldsymbol{\delta}_\theta|^q] \le 2^{q-1}\frac{1}{\pi\sqrt{2}}\Gamma\left(\frac{q+1}{2}\right)\left(\frac{\pi}{\sqrt{2}}\|\theta-\theta^\star\|\right)^{q+1} + 2^{q-2}\|\theta-\theta^\star\|^{2q}\,\mathbb{E}[p(-|\mathbf{z}|)|\mathbf{z}|^q] \\ + 2^{2(q-1)}\|\theta-\theta^\star\|^q\left(1-\frac{1}{4}\|\theta-\theta^\star\|^2\right)^{q/2}\frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right)\mathbb{E}[p(-|\mathbf{z}|)].\end{aligned}$$

**Proof.** Using the identity $[x]_+ = (|x| + x)/2$, we have for any $\theta \in S^{d-1}$,

$$\begin{aligned}\boldsymbol{\delta}_\theta = [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta]_+ - [-\mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star]_+ &= \frac{|\mathbf{x}^\mathsf{T}\theta| - \mathbf{y}\mathbf{x}^\mathsf{T}\theta}{2} - \frac{\operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\mathbf{x}^\mathsf{T}\theta^\star - \mathbf{y}\mathbf{x}^\mathsf{T}\theta^\star}{2} \\ &= \frac{(\operatorname{sign}(\mathbf{x}^\mathsf{T}\theta) - \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star))\mathbf{x}^\mathsf{T}\theta}{2} - \frac{(\mathbf{y} - \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star))\mathbf{x}^\mathsf{T}(\theta - \theta^\star)}{2}.\end{aligned}$$

Therefore, for any $q \geq 2$, Jensen's inequality implies

$$\begin{aligned}|\boldsymbol{\delta}_\theta|^q &\leq 2^{q-1}\left(\left|\frac{(\operatorname{sign}(\mathbf{x}^\mathsf{T}\theta) - \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star))\mathbf{x}^\mathsf{T}\theta}{2}\right|^q + \left|\frac{(\mathbf{y} - \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star))\mathbf{x}^\mathsf{T}(\theta - \theta^\star)}{2}\right|^q\right) \\ &= 2^{q-1}\mathbb{1}\{\operatorname{sign}(\mathbf{x}^\mathsf{T}\theta) \neq \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\}|\mathbf{x}^\mathsf{T}\theta|^q + 2^{q-1}\mathbb{1}\{\mathbf{y} \neq \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\}|\mathbf{x}^\mathsf{T}(\theta - \theta^\star)|^q.\end{aligned}$$

Kuchelmeister and van de Geer (2023, Corollary A.2.1) showed that

$$\begin{aligned}\mathbb{E}[\mathbb{1}\{\operatorname{sign}(\mathbf{x}^\mathsf{T}\theta) \neq \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\}|\mathbf{x}^\mathsf{T}\theta|^q] &\leq \frac{1}{\pi\sqrt{2}}\frac{\Gamma\left(\frac{q}{2}+1\right)}{q+1}\left(\frac{\pi}{\sqrt{2}}\|\theta - \theta^\star\|\right)^{q+1} \\ &\leq \frac{1}{\pi\sqrt{2}}\Gamma\left(\frac{q+1}{2}\right)\left(\frac{\pi}{\sqrt{2}}\|\theta - \theta^\star\|\right)^{q+1}\end{aligned}$$

where the latter inequality uses $\Gamma(q/2+1)/(q+1) \leq \Gamma((q+1)/2)$.

Using the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$,

$$\begin{aligned}\mathbb{E}[\mathbb{1}\{\mathbf{y} \neq \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\}|\mathbf{x}^\mathsf{T}(\theta - \theta^\star)|^q] &= \mathbb{E}[p(-|\mathbf{x}^\mathsf{T}\theta^\star|)|\mathbf{x}^\mathsf{T}(\theta - \theta^\star)|^q] \\ &= \mathbb{E}\left[p(-|\mathbf{z}|)|(\rho - 1)\mathbf{z} + \sqrt{1-\rho^2}\mathbf{z}_\perp|^q\right]\end{aligned}$$

where $\rho = \theta^\mathsf{T}\theta^\star$, and $\mathbf{z}$ and $\mathbf{z}_\perp$ are independent standard normal random variables. By Jensen's inequality,

$$|(\rho - 1)\mathbf{z} + \sqrt{1-\rho^2}\mathbf{z}_\perp|^q \leq 2^{q-1}|\rho - 1|^q|\mathbf{z}|^q + 2^{q-1}(1-\rho^2)^{q/2}|\mathbf{z}_\perp|^q.$$

Moreover, we have

$$1 - \rho = \frac{1}{2}\|\theta - \theta^\star\|^2 \quad \text{and} \quad 1 - \rho^2 = \|\theta - \theta^\star\|^2\left(1 - \frac{1}{4}\|\theta - \theta^\star\|^2\right).$$

Therefore, using independence of $\mathbf{z}$ and $\mathbf{z}_\perp$,

$$\begin{aligned}&\mathbb{E}[\mathbb{1}\{\mathbf{y} \neq \operatorname{sign}(\mathbf{x}^\mathsf{T}\theta^\star)\}|\mathbf{x}^\mathsf{T}(\theta - \theta^\star)|^q] \\ &\leq 2^{q-1}|\rho - 1|^q\,\mathbb{E}[p(-|\mathbf{z}|)|\mathbf{z}|^q] + 2^{q-1}(1-\rho^2)^{q/2}\,\mathbb{E}[p(-|\mathbf{z}|)]\,\mathbb{E}[|\mathbf{z}_\perp|^q] \\ &= \frac{1}{2}\|\theta - \theta^\star\|^{2q}\,\mathbb{E}[p(-|\mathbf{z}|)|\mathbf{z}|^q] + 2^{q-1}\|\theta - \theta^\star\|^q\left(1 - \frac{1}{4}\|\theta - \theta^\star\|^2\right)^{q/2}\frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right)\mathbb{E}[p(-|\mathbf{z}|)].\end{aligned}$$

■

**Proof of Lemma 8.** We use Lemma 20 with $p(t) = g'(\beta t)$. Therefore we need to bound $\mathbb{E}[g'(-\beta|\mathbf{z}|)]$ and $\mathbb{E}[g'(-\beta|\mathbf{z}|)|\mathbf{z}|^q]$ for all integers $q \geq 2$. Since $g'(-\beta|z|) \leq \exp(-\beta|z|)$ for

all $z \in \mathbb{R}$, we can use Lemma 16 for the former and Lemma 18 for the latter. We obtain

$$\begin{aligned}
\mathbb{E}[|\boldsymbol{\delta}_\theta|^q] &\le 2^{q-1} \cdot \frac{1}{\pi\sqrt{2}}\Gamma\left(\frac{q+1}{2}\right)\left(\frac{\pi}{\sqrt{2}}\|\theta-\theta^\star\|\right)^{q+1} \\
&\quad + 2^{q-2} \cdot \|\theta-\theta^\star\|^{2q} \cdot \frac{q!}{\beta^q} \\
&\quad + 2^{2(q-1)} \cdot \|\theta-\theta^\star\|^q \left(1-\frac{1}{4}\|\theta-\theta^\star\|^2\right)^{q/2} \frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right) \cdot \frac{1}{\beta}\sqrt{\frac{2}{\pi}} \\
&\le 2^{q-1} \cdot \frac{1}{\pi\sqrt{2}}\Gamma\left(\frac{q+1}{2}\right)\left(\frac{\pi}{\sqrt{2}}\|\theta-\theta^\star\|\right)^{q+1} + 2^{2q-2} \cdot \|\theta-\theta^\star\|^q \cdot \frac{q!}{\beta} \\
&\quad + 2^{2(q-1)} \cdot \|\theta-\theta^\star\|^q \frac{2^{q/2}}{\sqrt{\pi}}\Gamma\left(\frac{q+1}{2}\right) \cdot \frac{1}{\beta}\sqrt{\frac{2}{\pi}} \\
&\le q! \cdot (C\|\theta-\theta^\star\|)^{q-2} \cdot \frac{C}{2} \cdot \|\theta-\theta^\star\|^2 \cdot \left(\|\theta-\theta^\star\| + \frac{1}{\beta}\right)
\end{aligned}$$

for some absolute constant $C > 0$, where the final inequality uses $\Gamma((q+1)/2) \le q!$. The final right-hand side in the previous display is clearly bounded above by $q! v b^{q-2}/2$ for the specified choices of $b$ and $v$. ∎

## Appendix C. Proof of Lemma 14

**Proof of Lemma 14.** We have

$$\begin{aligned}
\mathrm{err}_{\theta^\star}(\theta) &= \mathbb{E}\big[g'(-\beta\,\mathrm{sign}(\mathbf{x}^\mathsf{T}\theta)\mathbf{x}^\mathsf{T}\theta^\star)\big] \\
&= \mathbb{E}\big[g'(-\beta\,\mathrm{sign}(\mathbf{x}^\mathsf{T}\theta)\mathbf{x}^\mathsf{T}\theta^\star)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0\}\big] + \mathbb{E}\big[g'(-\beta\,\mathrm{sign}(\mathbf{x}^\mathsf{T}\theta)\mathbf{x}^\mathsf{T}\theta^\star)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star > 0\}\big] \\
&= \mathbb{E}\big[(1-g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|))\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0\}\big] + \mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star > 0\}\big] \\
&= \Pr(\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0) - 2\,\mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0\}\big] + \mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\big] \\
&= \Pr(\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0) - 2\,\mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0\}\big] + \mathrm{err}_{\theta^\star}(\theta^\star).
\end{aligned}$$

Therefore

$$\begin{aligned}
\Pr(\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0) &= \mathrm{err}_{\theta^\star}(\theta) - \mathrm{err}_{\theta^\star}(\theta^\star) + 2\,\mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\mathbb{1}\{\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0\}\big] \\
&\le \mathrm{err}_{\theta^\star}(\theta) - \mathrm{err}_{\theta^\star}(\theta^\star) + 2\,\mathbb{E}\big[g'(-\beta|\mathbf{x}^\mathsf{T}\theta^\star|)\big].
\end{aligned}$$

The claim now follows by Lemma 17 and the fact that

$$\Pr(\theta^\mathsf{T}\mathbf{x}\mathbf{x}^\mathsf{T}\theta^\star \le 0) = \frac{\arccos(\theta^\mathsf{T}\theta^\star)}{\pi} \ge \frac{\sqrt{2(1-\theta^\mathsf{T}\theta^\star)}}{\pi} = \frac{\|\theta-\theta^\star\|}{\pi}.$$

∎