Constants Matter: The Performance Gains of Active Learning

Stephen Mussmann [1] **Sanjoy Dasgupta** [2]

Abstract

Within machine learning, active learning studies the gains in performance made possible by adaptively selecting data points to label. In this work, we show through upper and lower bounds, that for a simple benign setting of well-specified logistic regression on a uniform distribution over a sphere, the expected excess error of both active learning and random sampling have the same inverse proportional dependence on the number of samples. Importantly, due to the nature of lower bounds, any more general setting does not allow a better dependence on the number of samples. Additionally, we show a variant of uncertainty sampling can achieve a faster rate of convergence than random sampling by a factor of the Bayes error, a recent empirical observation made by other work. Qualitatively, this work is pessimistic with respect to the asymptotic dependence on the number of samples, but optimistic with respect to finding performance gains in the constants.

1. Introduction

Given samples of input-label pairs, machine learning algorithms return decision rules that will predict future labels given inputs. Active learning studies the possible reduction in error if the samples are adaptively chosen by the machine learning system rather than randomly sampled. Active learning algorithms have been demonstrated to reduce error in a variety of both theoretical and empirical settings.

Theoretically, there are a variety of cases where the excess error of active learning has a better dependence (polynomially or even exponentially) on the number of samples, n, than the excess error of random sampling (Balcan et al., 2007; Balcan & Long, 2013; Wang & Singh, 2016). In this work, we show that for a simple benign setting of well-specified logistic regression on a uniform distribution over a sphere, the expected excess error of both active learning and random sampling have the same inverse proportional dependence on n: the expected excess error decreases as $\Theta(1/n)$.

Table 1. A table with our lower bounds (impossibility results) and our upper bounds (algorithm analysis) for both random sampling and adaptive sampling (active learning) in our setting. c and c' are universal constants and err^* is the Bayes error.

Sampling	Result type	Exp. Excess Error
Adaptive	Impossibility	$\geq c\ \text{err}^* \frac{d}{n}$
Adaptive	Alg. Analysis	$\leq c' \text{err}^* \frac{d}{n}$
Random	Impossibility	$\geq c\ \frac{d}{n}$
Random	Alg. Analysis	$\leq c' \frac{d \log d}{n}$

On a more optimistic note, we show that active learning can reduce the expected excess error by a distribution-dependent factor, the Bayes error, which we denote err^*. As our setting employs well-specified logistic regression as the label distribution, this result matches existing empirical observations for logistic regression (Mussmann & Liang, 2018a).

A list of the results is shown in Table 1. Note that all results require the dimension to be larger than a constant and the number of samples to be sufficiently large in terms of the specification of the setting (the dimension d, the radius of the input distribution sphere r, and the norm of the true parameters M). Finally, both upper bounds require the Bayes error, err^*, to be smaller than a constant. Importantly, because of the simplicity of our setting, the lower bounds are quite strong while the upper bounds are quite weak and only show the (almost) tightness of the lower bounds.

The two lower bounds are proved using a variant of Fano's inequality (Duchi & Wainwright, 2013; Scarlett & Cevher, 2019) and a carefully designed set of possible logistic regression weights. The proofs use very similar arguments that differ in the bound on the mutual information. The random sampling upper bound is shown for the maximum likelihood estimator (MLE). While the MLE has been analyzed many times (Van der Vaart, 2000; Lehmann & Casella, 2006; Frostig et al., 2015), our result differs in that we prove

[1]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, USA [2]Computer Science & Engineering, University of California, San Diego, USA. Correspondence to: Stephen Mussmann <mussmann@cs.washington.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

the upper bound in terms of the excess zero-one loss, not the excess logistic loss, a non-trivial difference. Finally, the adaptive algorithm analyzed for the upper bound is a two-step algorithm where the first half of the queries are randomly sampled and the second half of the queries are collected by uncertainty sampling, and used for gradient updates; that is, instead of optimizing the logistic loss on all labels thus far, we only take a gradient step at the most recently labeled point. For the analysis, we use the insight that uncertainty sampling is roughly stochastic gradient descent on the zero-one loss (Mussmann & Liang, 2018b) and then adapt a standard stochastic gradient descent convergence argument (Rakhlin et al., 2012). Finally, we present illustrative synthetic experimental results for our upper bounds, demonstrating the effect of the problem dependent parameters in our setting.

In summary, our contributions are threefold:

- Presentation of four results suggesting that the advantage of active learning is often not in the dependence on the number of samples but in the constants, particularly the Bayes error.
- Complete, self-contained proofs of all results from basic principles with the exception of Fano's inequality and a few concentration results from other works.
- Synthetic experiments for the logistic regression uniform sphere setting to illustrate our two upper bounds.

The paper is organized as follows: we first review related work in Section 2 before introducing our notation and setting in Section 3. We then proceed with our results in Section 4, present synthetic experiments in Section 5, discuss some implications in Section 6, and conclude with Section 7.

2. Related work

The problem of learning homogeneous linear separators over data drawn from the uniform distribution on a sphere has been a fruitful setting for active learning (Dasgupta et al., 2005; Dasgupta, 2005; Balcan et al., 2007; 2009; Wang & Singh, 2016). In the realizable case, also known as the noiseless case, it is known that active learning enables dramatic performance gains. While random sampling methods require $\Theta(1/\varepsilon)$ samples to achieve ε error, a variety of active learning methods achieve a sample complexity of $\Theta(\ln(1/\varepsilon))$. This improvement is referred to as exponential because the error rate goes from $\Theta(1/n)$ to $\exp(-\Theta(n))$.

In the presence of general noise, exponential gains are not possible (Kääriäinen, 2006; Beygelzimer et al., 2009). In particular, the sample complexity for general active learning grows as $\Theta(1/\varepsilon^2)$ (the excess error is $\Theta(1/\sqrt{n})$) which is the same dependence on ε as general passive learning. Interestingly, the sample complexity lower bounds between passive and active learning differ by a factor of the minimal error (Hanneke, 2014), a quantity similar to err^*.

The Tsybakov noise condition (see assumption (A1) in Tsybakov (2004)) is an important quantity for characterizing sample complexities. Intuitively, the Tsybakov noise condition relates how the excess error of a classifier scales with the disagreement between the classifier and the optimal classifier. Using the notation of Tsybakov (2004), this scaling is measured by a variable $\kappa \geq 1$ which is a function of the true data distribution and a set of classifiers $\mathcal{F}$:

$$\begin{aligned}&\exists c, c' > 0 \text{ such that } \forall f, f' \in \mathcal{F}: \\ &\Pr(f(x) \neq f'(x)) \leq c \implies \\ &\mathbb{E}_x[\mathbf{1}[f(x) \neq f'(x)]|\Pr(Y=1|x) - \Pr(Y=0|x)|] \\ &\geq c' \Pr_x(f(x) \neq f'(x))^\kappa\end{aligned}$$

Perhaps the two most common cases are $\kappa = 1$ and $\kappa = 2$. $\kappa = 1$ implies noise that is easier to handle; examples include noiseless (deterministic) label distributions and cases where the conditional label distribution is bounded away from $1/2$ (Massart noise (Massart & Nédélec, 2006)). $\kappa = 2$ involves more noise near the decision boundary. For example, if the conditional label distribution behaves linearly with non-zero slope across the decision boundary, as is the case with logistic regression, then $\kappa = 2$.

The Tsybakov noise condition importantly separates exponential and polynomial error rates, and for polynomial rates, determines the polynomial exponent. For example, Balcan et al. (2007) provide an analysis of an uncertainty sampling variant (known as margin sampling) for homogeneous linear classifiers on the uniform distribution over a sphere and show a key dependence on the Tsybakov noise condition. In the notation of Balcan et al. (2007), $\alpha = 1 - 1/\kappa$. In particular, if $\alpha = 0$ ($\kappa = 1$), the excess error drops exponentially fast in the number of samples. However, for $\alpha \in (0, 1)$ ($\kappa > 1$), the rate is significantly slower, the excess error decays polynomially in the number of samples. In particular, for $\alpha = 1/2$ ($\kappa = 2$), the excess error goes as $\tilde{\mathcal{O}}(1/n)$, approximately (up to log terms) the error rate we find for logistic regression. Balcan & Long (2013) gives similar results for log-concave distributions.

The most similar lower bounds to ours in the literature are in terms of the Tsybakov noise condition. Castro & Nowak (2007) shows that for distributions where the decision boundary is smooth and $\kappa = 2$, the active learning excess error rate is $\Omega(1/n)$. Hanneke & Yang (2015) shows the same for distributions where $\kappa = 2$ in terms of constants such as those from the definition of the Tsybakov noise condition. Most recently, Wang & Singh (2016) gave a lower bound of $\Omega(1/n)$ with a construction of a label distribution with Tsybakov noise condition $\kappa = 2$ and the input distribu-

tion uniform on a sphere. In contrast to the previous lower bounds, we incorporate the following two aspects:

- We prove a stronger lower bound: rather than showing there is a hard-to-learn family of distributions within all distributions with a Tsybakov noise condition of $\kappa = 2$ (with an input distribution uniform on a sphere), we show that there is a hard-to-learn family of distributions within the strictly smaller set of well-specified logistic regression label distributions and inputs uniform on a sphere.

- We work out all constants in terms of intuitive problem setting quantities (dimension, Bayes error, radius of sphere, etc) and hide nothing in asymptotic notation to importantly bring light to the gains possible in the constants, rather than the dependence on the number of samples.

Interestingly, while general random sampling lower bounds for Tsybakov noise condition $\kappa = 2$ yield a $\Omega(1/n^{2/3})$ rate, we show that a random sampling algorithm yields the standard $O(1/n)$ error rate for logistic regression.

3. Setting

We study a binary classification setting with an input set $\mathcal{X}$ and a label set $\mathcal{Y}$. In particular, we set $\mathcal{X} \subset \mathbb{R}^d$ and $|\mathcal{Y}| = 2$. The goal is to identify a measurable $f : \mathcal{X} \to \mathcal{Y}$ that achieves low error, where the error is defined as

$$\text{err}(f) = \Pr(f(x) \neq y) \tag{1}$$

In this work, we analyze the case that the input data distribution is uniform on a radius r sphere: $\mathcal{X} = rS^{d-1} = \{x \in \mathbb{R}^d : \|x\| = r\}$. Throughout the paper, we require $d \geq 5$, and for the lower bounds we require $d \geq 24$.

We further assume the true label distribution is well-specified logistic regression with parameters w^* of norm M. Specifically,

$$\Pr(y|x) = \sigma(yx \cdot w^*) \tag{2}$$

where $\|w^*\| = M$, $y \in \{-1, 1\}$, and σ is the standard sigmoid function: $\sigma(u) = 1/(1 + \exp(-u))$.

Because the Bayes-optimal classifier is linear for well-specified logistic regression, we are especially interested in linear classifiers:

$$f(x; w) = \begin{cases} -1 & w \cdot x < 0 \\ 1 & w \cdot x \geq 0 \end{cases} \tag{3}$$

We define the error of weights w as

$$\text{err}(w) = \Pr(f(x; w) \neq y) \tag{4}$$

and the Bayes error as

$$\text{err}^* = \text{err}(w^*) \tag{5}$$

Note that by spherical symmetry, err^* only depends on the norm of w^*, that is M, and not on the direction of w^*.

We can think of three variables defining the setting: the norm of the true parameters M, the radius r of the sphere from which points are drawn, and the dimension d of the input space. Although it appears there are three variables, there are effectively only two. Note that since the true label distribution depends only on $x \cdot w^*$, if we double M and halve r, the structure of the setting stays the same. Therefore, we can think of the product Mr and the dimension d as the two defining variables.

Furthermore, given a fixed d, there is a bijection between $Mr \in [0, \infty)$ and $\text{err}^* \in (0, 1/2]$. So we can also think of parametrizing settings by err^* and d.

We consider two methods of data collection: adaptive sampling and random sampling. A random sampling algorithm is given n points sampled from the data distribution (inputs uniform on $\mathcal{X}$ with labels according to well-specified logistic regression) and returns a classifier $\hat{f}$. An adaptive sampling algorithm iteratively selects a point in $\mathcal{X}$ and receives a label drawn from the true conditional data distribution. After this process is repeated n times (n labels have been observed), the algorithm returns a classifier $\hat{f}$. Importantly, an adaptive sampling algorithm's selection of a point can depend on previously observed labels.

4. Results

In this work, we show four results: upper and lower bounds for the adaptive and random sampling settings. We begin by presenting a few lemmas that provide intuition for the setting.

4.1. Bayes' error

First, we analyze the Bayes error and show that it is linearly related to $\frac{\sqrt{d-1}}{Mr}$. Intuitively, for large d, any component of

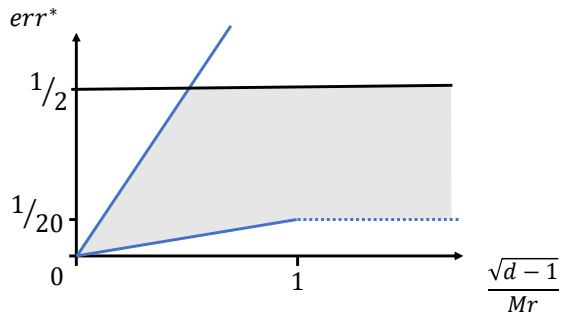

Figure 1. A graphical representation of the two bounds on the Bayes error err* in terms of the quantity $\frac{\sqrt{d-1}}{Mr}$. Each setting of M, r, and d corresponds to a point that must lie in the gray shaded region. Trivially, the Bayes error is below $1/2$. The upper blue line corresponds to Lemma 4.2 while the lower truncated blue line corresponds to Lemma 4.1. The dashed blue line is a consequence of the monotonicity of the Bayes error as a function of M.

a random point on a sphere will be approximately a normal distribution with variance $\Theta(r^2/d)$ and thus standard deviation $\Theta(r/\sqrt{d})$. Since w^* has norm M, $w^* \cdot x$ will be approximately normal with standard deviation $\Theta(Mr/\sqrt{d})$. The Bayes optimal classifier is unlikely to err when $|w^* \cdot x|$ is large, but when $w^* \cdot x$ is zero, the Bayes optimal classifier errs with probability $1/2$. Roughly speaking, the probability that $w^* \cdot x$ (normally distributed) is close to zero is inversely proportional to the standard deviation, $\Theta(Mr/\sqrt{d})$. Thus, we might expect that the Bayes error scales as $\Theta(\sqrt{d}/Mr)$. In fact, this is the case as seen in the following two lemmas.

Lemma 4.1. *Suppose $d \geq 5$. If $\frac{\sqrt{d-1}}{Mr} \leq 1$, then,*

$$err^* \geq \frac{1}{20}\frac{\sqrt{d-1}}{Mr} \tag{6}$$

Lemma 4.2. *Suppose $d \geq 5$.*

$$err^* \leq \frac{8}{7}\frac{\sqrt{d-1}}{Mr} \tag{7}$$

Proofs for these two lemmas can be found in Appendix A. The implications of these two lemmas are shown graphically in Figure 1. We see that if err* is sufficiently small (less than $1/20$), err* scales linearly with $\frac{\sqrt{d-1}}{Mr}$.

4.2. Excess error and angle

We now relate the excess error of an estimated hypothesis $\hat{w}$ to the angle between $\hat{w}$ and w^*: $\angle(\hat{w}, w^*)$. Note that the error of a decision rule defined by weights w does not depend on the norm $\|w\|$, but only on the direction of w, and by spherical symmetry, only depends on the angle $\angle(w, w^*)$. In our settings, the angular estimation error is proportional to the disagreement between the decision rules defined by $\hat{w}$ and w^*. Thus, the scaling of the excess error as a function of the angle is closely related to the Tsybakov noise condition. We find that the excess error scales as the square of the angle as shown in the following lemma, yielding a Tsybakov noise condition of $\kappa = 2$ for our setting.

Lemma 4.3. *Suppose $d \geq 5$. For any w such that $\angle(w, w^*) \leq \frac{\pi}{2}$,*

$$err(w) - err(w^*) \leq \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\angle(w, w^*)^2. \tag{8}$$

Furthermore, for any w such that $Mr\angle(w, w^*) \leq 1$ and $\angle(w, w^*) \leq \frac{\pi}{2}$,

$$err(w) - err(w^*) \geq \frac{1}{25}\frac{Mr}{\sqrt{d-1}}\angle(w, w^*)^2. \tag{9}$$

A proof for this lemma can be found in Appendix A. To gain intuition for this quadratic dependence, see Figure 2.

4.3. Fano's inequality and mutual information tools

We now discuss some concepts used in the proofs of our lower bounds. We wish to show that the expected excess error is bounded below.

Fano's inequality (Fano, 1961) is an information-theoretic tool often used for proving impossibility results (Scarlett & Cevher, 2019). In this work, we make use of Fano's inequality to prove our two lower bounds (see Section 4.4). Here, we summarize some of the tools that can be found in the survey Scarlett & Cevher (2019).

The setup of the standard Fano's inequality is the following: there is an unknown random variable of interest, V^*, which is drawn uniformly at random from a finite set $\mathcal{V}$ and an estimate of V^* known as $\hat{V}$ also in the finite set $\mathcal{V}$.

Fano's inequality is a mathematical identity used to upper bound the probability that the estimator is correct based on the mutual information between $\hat{V}$ and V^*:

$$\Pr(\hat{V} = V^*) \leq \frac{I(V^*; \hat{V}) + \log 2}{\log |\mathcal{V}|} \tag{10}$$

Thus, if we can upper bound the mutual information between $\hat{V}$ and V^* for any estimator $\hat{V}$, we can upper bound the accuracy of the estimator. However, because $\hat{V}$ can take many forms, how do we bound the mutual information?

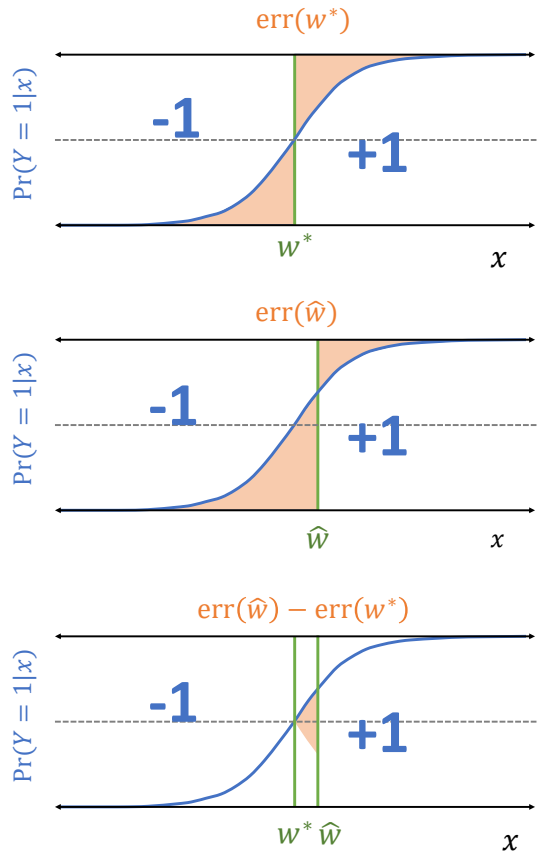

Figure 2. A diagram for intuition of the quadratic scaling of the excess error with the angle. This diagram shows the related setting of threshold logistic regression. The x-axis is the input dimension while the y-axis is the probability of observing a positive label. The blue curve is a logistic curve. In the top pane, we see the threshold decision rule w^* represented by the vertical green line. The area of the orange shaded region corresponds to the error of w^*. In the middle pane we see an estimate $\hat{w}$ and the error of $\hat{w}$ as the area of the orange shaded region. Finally, in the bottom pane, we see a diagram of the excess error taken by subtracting the top pane from the middle pane. A rearrangement yields that the excess error corresponds to the area of a triangle with width and height determined by the difference between w^* and $\hat{w}$. Thus, we see that the excess error scales as the square of the difference between w^* and $\hat{w}$. A similar, but harder to visualize, scenario occurs for logistic regression on the sphere.

An important tool is the data processing inequality (Cover & Thomas, 1999). A common situation in machine learning includes the following three random variables: the true parameters V^*, some data $\{(X_i, Y_i)\}_{i=1}^n$, and the estimate $\hat{V}$. Here, we assume that $\hat{V}$ and V^* are conditionally independent given the data $\{(X_i, Y_i)\}_{i=1}^n$; in other words, the estimate $\hat{V}$ doesn't depend on the true parameters V^* except through the data generated by V^*. Under this condition,

$$I(V^*; \hat{V}) \leq I(V^*; \{(X_i, Y_i)\}_{i=1}^n) \tag{11}$$

Intuitively, this means we can not create information by post-processing: the estimate $\hat{V}$ can't be more informative than the raw data $\{(X_i, Y_i)\}_{i=1}^n$. However, we still need a technique to bound the mutual information between the raw observations and the true parameters.

In the adaptive setting, there is a complex dependency between the data points. Specifically, we generate X_i based on past observations $\{(X_j, Y_j)\}_{j=1}^{i-1}$ then we observe Y_i based on the true parameters V^* and X_i. Fortunately, a tool known as tensorization (Scarlett & Cevher, 2019), based on the chain rule for mutual information (Cover & Thomas, 1999), yields the following,

$$I(V^*; \{(X_i, Y_i)\}_{i=1}^n) \leq \sum_{i=1}^n I(V^*; Y_i | X_i) \tag{12}$$

Altogether, we can upper bound the probability that any estimator is correct if we can upper bound the sum of the information "leaked" by each of the observations.

4.4. Lower bounds

Both lower bounds are proved using a construction of 2^{d-1} hard-to-distinguish label distributions parametrized by $\varepsilon > 0$. Intutively, ε corresponds to the amount of separation between the distributions. Define

$$\mathcal{W} = \frac{M}{\sqrt{1 + (d-1)\varepsilon^2}}(1, \pm\varepsilon, \pm\varepsilon, \ldots, \pm\varepsilon) \tag{13}$$

Note that all elements of $\mathcal{W}$ have the same norm of M. We define a set of 2^{d-1} conditional label distributions $\Pi = \{\pi_w : w \in \mathcal{W}\}$ where π_w is the conditional label probability for logistic regression with weights w: $\pi_w(x) = \Pr(Y = 1 | X = x; w) = \sigma(yx \cdot w)$. We let the true conditional label distribution π^* be drawn uniformly from Π, or equivalently, w^* is drawn uniformly from $\mathcal{W}$. If we can bound the mutual information between π^* and Y_i

conditioned on X_i, we can use Fano's inequality, the data processing inequality, and tensorization to show it is impossible for an estimator in Π to estimate π^* well with high probability. However, what about decision rules? Maybe we can do better if we return a decision rule that isn't even linear?

We can handle these concerns via a reduction. Let $\rho(f, \pi) = \Pr_{x,y\sim\pi(x)}(y \neq f(x))$, in other words, the error of the decision rule f under the conditional label distribution π. For any estimated decision rule $\hat{f}$, define

$$\hat{\pi} = \operatorname*{argmin}_{\pi\in\Pi} \rho(\hat{f}, \pi) \tag{14}$$

The proof becomes somewhat technical at this point, but one can note that if $\hat{f}$ has sufficiently low error, then $\hat{\pi} = \pi^*$. We can additionally use the data processing inequality a second time:

$$I(\pi^*; \hat{\pi}) \leq I(\pi^*; \hat{f}) \tag{15}$$

This discussion of the lower bounds has left out a quite important detail: our proofs use an approximate recovery form of Fano's inequality. Instead of upper bounding the probability that $\hat{\pi} = \pi^*$, we define an appropriate similarity relation so that we can upper bound the probability that $\hat{\pi}$ and π^* are similar (Duchi & Wainwright, 2013; Scarlett & Cevher, 2019). Without this detail, we would lose the factor of d in our lower bounds.

Theorem 4.4. *Fix $d \geq 24$, $M > 0$, $r > 0$. For sufficiently large n (in terms of M, r, d, and err^*), for any data collection strategy for n data points and estimator $\hat{f}$ depending on those data points (and conditionally independent of the true label distribution given the data), there exists a norm-M w^* such that,*

$$\mathbb{E}[err(\hat{f})] - err^* \geq \frac{1}{250000} err^* \frac{d}{n}. \tag{16}$$

Theorem 4.5. *Fix $d \geq 24$, $M > 0$, $r > 0$. For sufficiently large n (in terms of M, r, d, and err^*), for any estimator $\hat{f}$ computed from n random samples (and conditionally independent of the true label distribution given the data), there exists a norm-M w^* such that:*

$$\mathbb{E}[err(\hat{f})] - err^* \geq \frac{1}{22000000} \frac{d}{n}. \tag{17}$$

The proofs for these theorems are in Appendix C and Appendix D, respectively. The lemmas for the two arguments are very similar and can be found in Appendix B. We have hidden how large n must be in both theorems, but full conditions on n can be found in the statements of the theorems in the appendix.

Note that we have not hidden any constants. The universal constants in the front are rather small. Note that we did not optimize the universal constants and they are most likely very loose.

4.5. Algorithms for upper bounds

In this section, we define the random sampling and adaptive sampling algorithms used for our upper bounds. Briefly, the random sampling algorithm is simply logistic regression maximum likelihood estimation and the adaptive algorithm is an uncertainty sampling variant: random sample with half of the budget n to initialize the weights, then use the other half of the budget to iteratively take uncertainty sampling gradient steps with a decaying step size.

4.5.1. Random sampling

Given n random samples $\{(x_i, y_i)\}_{i=1}^n$, the random sampling estimate is the weights that maximize the probability of observing the labels:

$$\hat{w} = \operatorname*{argmax}_w \Pr(\forall i : Y_i = y_i | w, \forall i : X_i = x_i) \tag{18}$$

$$= \operatorname*{argmax}_w \prod_{i=1}^n \sigma(y_i x_i \cdot w) \tag{19}$$

$$= \operatorname*{argmin}_w \frac{1}{n} \sum_{i=1}^n -\log \sigma(y_i x_i \cdot w) \tag{20}$$

$$\tag{21}$$

This motivates us to define the loss function $\ell(x, y, w)$ as the negative log-likelihood: $\ell(x, y, w) = -\log \sigma(y_i x_i \cdot w)$ and the empirical loss $\hat{L}_n(w) = \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, w)$ yielding $\hat{w} = \operatorname{argmin}_w \hat{L}_n(w)$.

Intuitively, if we define $L(w) = \mathbb{E}[\ell(x, y, w)]$, then $w^* = \operatorname{argmin}_w L(w)$ and $\hat{L}_n(w) \to_P L(w)$ as $n \to \infty$, so we might hope $\hat{w} \approx w^*$ for large n.

4.5.2. Adaptive sampling

For adaptive sampling, we begin with random sampling for the first $n/2$ samples. Then, an estimate of the true weights is calculated by minimizing the logistic loss, or equivalently maximizing the likelihood. We then rescale the estimate to produce w_1; this is done to ensure $\|w_1\| \leq M$ with high probability. A constraint set $\mathcal{W}$ is constructed as the intersection of a origin-centered sphere of radius $\|w_1\|$ and a cone around w_1. In the next phase, we proceed by iterations of uncertainty sampling gradient updates with an

L2 orthogonal projection onto $\mathcal{W}$.

For each iteration, we randomly draw x_t from the decision boundary defined by the current weight iterate w_t: $\{x \in \mathcal{X} : w_t \cdot x = 0\}$. After querying the label of x_t as y_t, we compute the gradient $g_t = \nabla_w \ell(x_t, y_t, w_t)$ on the new point. Finally, we update $w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t g_t)$ where $\eta_t = \frac{1}{\lambda t}$ and $\lambda = \frac{1}{12}\frac{r^2}{d-1}$. After $n/2$ iterations, we have exhausted the label budget, and we return the last iterate. This entire process is shown as Algorithm 1. Note that the algorithm does not require knowledge of M or err*.

Algorithm 1 Active learning algorithm

Set convexity parameter $\lambda = \frac{r^2}{12(d-1)}$
Randomly sample and label $n/2$ points as B_{random}
Compute $w_{\text{random}} = \text{argmin}_w \sum_{(x,y)\in B_{\text{random}}} \ell(x, y, w)$ for $\ell(x, y, w) = -\log \sigma(yx \cdot w)$
Set $w_1 = 2w_{\text{random}}/3$
Set $\theta_{\max} = \frac{1}{2}\min\left(\frac{\pi}{2}, \frac{2}{3\|w_1\|r}\right)$
Set $\mathcal{W} = \{w : \|w\| \le \|w_1\|, \angle(w, w_1) \le \theta_{\max}\}$
for $t = 1, \ldots, n/2$ **do**
  Sample x_t uniformly from $\{x : \|x\| = r, x \cdot w_t = 0\}$
  Label x_t to get y_t
  Compute $g_t = \nabla_w \ell(x_t, y_t, w_t) = -\frac{1}{2} y_t x_t$
  Compute $w_{t+1} = \Pi_{\mathcal{W}}(w_t - \eta_t g_t)$ where $\eta_t = \frac{1}{\lambda t}$
end for
Return: $\hat{w} = w_{n/2+1}$

4.6. Upper bounds

In this section, we present the guarantees for the random sampling and adaptive sampling algorithms. The random sampling guarantee follows from a second-order Taylor expansion along with several geometric lemmas. The adaptive sampling guarantee uses the random sampling guarantee for the initialization, then proceeds with a standard stochastic gradient descent analysis (Rakhlin et al., 2012; Nemirovski et al., 2009). The choice of this type of analysis is inspired by the observation that uncertainty sampling roughly corresponds to stochastic gradient descent steps on the zero-one loss (Mussmann & Liang, 2018b).

Theorem 4.6. *Fix $d \ge 5$, $M > 0$, $r > 0$ such that $\frac{\sqrt{d-1}}{Mr} \le 1/12$. For sufficiently large n (in terms of M, r, d, and err*), for $\hat{w}$ as the logistic MLE estimator from n randomly sampled points,*

$$\mathbb{E}[err(\hat{w})] - err(w^*) \le 240000\frac{d\log(d)}{n}. \tag{22}$$

Theorem 4.7. *Fix $d \ge 5$, $M > 0$, $r > 0$ such that $\frac{\sqrt{d-1}}{Mr} \le 1/12$. For sufficiently large n (in terms of M, r, d, and err*), for the estimator $\hat{w}$ returned from Algorithm 1,*

$$\mathbb{E}[err(\hat{w})] - err(w^*) \le 26001 err^* \frac{d}{n}. \tag{23}$$

The proofs for these Theorems are in Appendix E and Appendix F, respectively. As with the lower bounds, full conditions on n can be found in the statements of the theorems in the appendix.

The universal constants in the front are rather large, though they they are most likely very loose from a lack of optimization. We can interpret the condition $\frac{\sqrt{d-1}}{Mr} \le 1/12$ with the lemmas from Section 4.1. For example, if err$^* \le 1/240$, the condition is satisfied.

5. Experiments

In this section, we run experiments in our synthetic setting: well-specified logistic regression with a uniform distribution over a radius r sphere. We compare random sampling and our adaptive algorithm (Algorithm 1) for varying n (Figure 3), M (Figure 4), and d (Figure 5). We fix $r = 1$ in all cases since learning behavior only depends on the product Mr. We make one small change to the algorithm and set $\theta_{\max} = \frac{\pi}{4}$ instead of $\theta_{\max} = \min(\pi/4, 1/(3\|w_1\|r))$. We found experimentally that the latter would require a larger n. All experiments are run with 100 replicates with error bars as 95% confidence intervals using a Gaussian approximation.

The test excess error of both random sampling and our adaptive algorithm appear to have an empirical inverse linear dependency on n (as predicted by our theory). Furthermore, while the test excess error of random sampling appears nearly independent of the limiting error, err*, the test excess error for our adaptive algorithm seems to scale linearly with err*. The dependence of random sampling's test excess error on the dimensionality d is less clear: it could be superlinear (as predicted by the upper bound) or it could be linear (matching the lower bound).

6. Discussion

Mussmann & Liang (2018a) show both experimentally and theoretically that the data efficiency of uncertainty sampling is inversely proportional to the limiting error, which is the same as the Bayes error for well-specified logistic regression. In that work, the data efficiency (of an active learning algorithm relative to random sampling) is defined as the ratio of the sample complexities, or the factor reduction in data samples required for an algorithm to match the performance of random sampling. Our four results allow us to compare lower and upper bounds on the expected excess error to compute both a lower and upper bound on the data efficiency. In particular, the data efficiency is at least

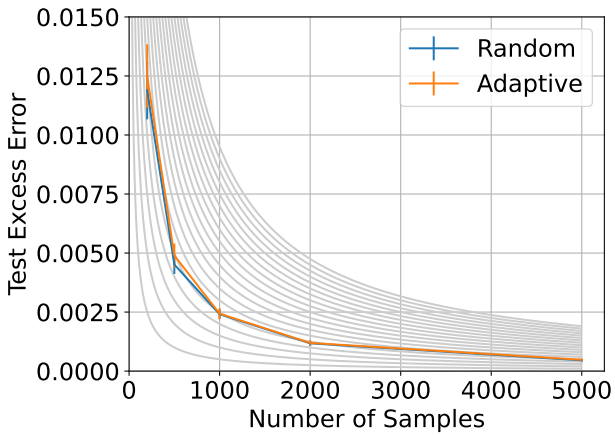

Figure 3. A plot comparing the test excess error of random sampling and our adaptive algorithm (Algorithm 1) for a varying number of samples n. The gray curves are of the form α/n to show the inverse dependence on n. We fix $r = 1$, $d = 10$, and $M = 20$, yielding between 7% and 8% limiting error.

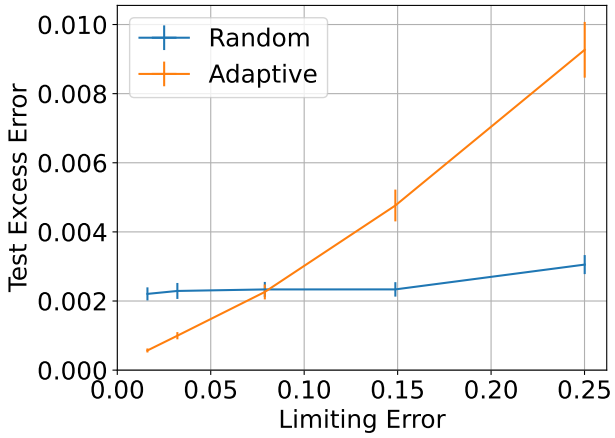

Figure 4. A plot comparing the test excess error of random sampling and our adaptive algorithm (Algorithm 1) for setups with varying the norm of the true parameters M (which changes the limiting error err^*). We fix $r = 1$, $d = 10$, and $n = 1000$ while varying $M \in \{5, 10, 20, 50, 100\}$. We note that while random sampling's test excess error remains approximately constant, adaptive sample's excess error grows linearly with the limiting error as our bounds show.

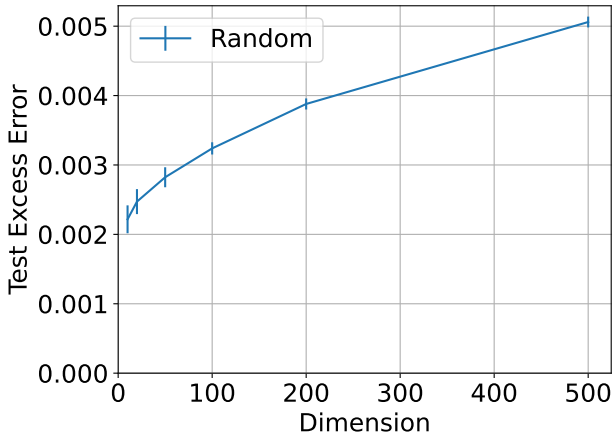

Figure 5. A plot showing the test excess error of the MLE performance of random sampling for various dimensionalities d. We set $r = 1$, $M = 20$, and $n = 100 \cdot d$. Note that although d/n remains constant, the test excess error seems to not be constant as a function of d, hinting that there may be an additional dependence on d, such as $\log(d)$. We did not run adaptive sampling for varying d because changing d affects the limiting error, err^*.

$\Omega(1/\text{err}^*)$ and is at most $O(\log(d)/\text{err}^*)$.

One may wonder what is the correct data efficiency for our setting. In the proof of the upper bound for random sampling, where the $\log d$ appears, we use two concentration inequalities: a Bernstein inequality applied dimension-by-dimension and a Matrix Chernoff inequality (Tropp, 2015). The union bound associated with the Bernstein inequality over d dimensions incurs a possibly unnecessary factor of $\log d$. Although Tropp (2015) show that their Matrix Chernoff bound is tight for the assumptions they make (see the coupon-collector discussion in Tropp (2015)), it seems that our setting is easier and so we may be able to shave off a $\log d$. In particular, the rank one matrices which compose the Hessian of the empirical loss will be nearly uniformly distributed in all directions, unlike the harder coupon-collector example. Thus, from a theoretical perspective, the $\log d$ is perhaps loose. However, the synthetic experiment may hint that the $\log d$ factor is necessary and perhaps the lower bound is loose. We note the possibility that random sampling with MLE includes $\log d$ while another estimator does not.

In this work we considered the adaptive sampling and random sampling settings. There are other important settings in between adaptive and random sampling. In particular, we note that the techniques in this work are insufficient to give (non-trivial) results for non-adaptive experimental design and batched adaptive sampling. Non-adaptive experimental design is the setting where the points to be labelled are chosen by the algorithm, similar to adaptive sampling, but

all labels are revealed at the same time so that adaptivity is not possible. Batched adaptive sampling is the setting where an algorithm makes queries in the form of b points to be labelled at the same time where a batch can be chosen based on the labels of previous batches. Note that the batched adaptive sampling setting is a generalization of adaptive sampling (with $b = 1$) and non-adaptive experimental design (with $b = n$). These settings form a nested hierarchy where random sampling is the most restrictive setting and adaptive sampling is the most powerful setting. A possible analysis strategy for batched adaptive sampling is an stochastic gradient descent convergence argument that makes use of the covariance of the gradient which decays as the batch size b grows; perhaps contributing a $\sqrt{b}$ factor to the expected excess error. A possible analysis strategy for non-adaptive experimental design is to dramatically expand $\mathcal{W}$, the set of possible w^*, or to apply some clever symmetry argument. We conjecture that the expected excess error rate of non-adaptive experimental design is the same as random sampling for our setting.

7. Conclusion

In summary, we analyzed upper and lower bounds on the expected excess error for both adaptive sampling and random sampling for a simple benign setting of well-specified logistic regression on inputs drawn uniformly at random from a sphere. Most importantly, all bounds had the same dependence on the number of samples. Because of the simplicity and naturalness of the construction for the lower bounds, this paper contributes evidence that in most practical cases, the advantage of active learning does not lie in a improved dependence on the number of samples, but rather in the problem dependent constants. As a result, we might abandon hope that active learning can provide gains for all problems, but instead search for problems where the constants are advantageous to active learning.

Acknowledgements

We thank the four anonymous reviewers whose comments and suggestions helped improve and clarify this manuscript. This work was supported by NSF CCF-1813160. SM was supported by NSF Graduate Fellowship DGE-1656518 and as an IFDS Postdoctoral Scholar with award NSF TRIPODS II-DMS 2023166.

References

Balcan, M.-F. and Long, P. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pp. 288–316. PMLR, 2013.

Balcan, M.-F., Broder, A., and Zhang, T. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 49–56, 2009.

Castro, R. M. and Nowak, R. D. Minimax bounds for active learning. In *International Conference on Computational Learning Theory*, pp. 5–19. Springer, 2007.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 1999.

Dasgupta, S. Coarse sample complexity bounds for active learning. In *NIPS*, volume 18, pp. 235–242, 2005.

Dasgupta, S., Kalai, A. T., and Monteleoni, C. Analysis of perceptron-based active learning. In *International conference on computational learning theory*, pp. 249–263. Springer, 2005.

Duchi, J. C. and Wainwright, M. J. Distance-based and continuum fano inequalities with applications to statistical estimation. *arXiv preprint arXiv:1311.2669*, 2013.

Fano, R. M. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.

Frostig, R., Ge, R., Kakade, S. M., and Sidford, A. Competing with the empirical risk minimizer in a single pass. In *Conference on learning theory*, pp. 728–763. PMLR, 2015.

Hanneke, S. Theory of active learning. *Foundations and Trends in Machine Learning*, 7(2-3), 2014.

Hanneke, S. and Yang, L. Minimax analysis of active learning. *J. Mach. Learn. Res.*, 16(12):3487–3602, 2015.

Kääriäinen, M. Active learning in the non-realizable case. In *International Conference on Algorithmic Learning Theory*, pp. 63–77. Springer, 2006.

Lehmann, E. L. and Casella, G. *Theory of point estimation*. Springer Science & Business Media, 2006.

Massart, P. and Nédélec, É. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Mussmann, S. and Liang, P. On the relationship between data efficiency and error for uncertainty sampling. In *International Conference on Machine Learning*, pp. 3674–3682. PMLR, 2018a.

Mussmann, S. and Liang, P. S. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. *Advances in Neural Information Processing Systems*, 31:6955–6964, 2018b.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Rakhlin, A., Shamir, O., and Sridharan, K. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1571–1578, 2012.

Scarlett, J. and Cevher, V. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8 (1-2):1–230, 2015.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Wang, Y. and Singh, A. Noise-adaptive margin-based active learning and lower bounds under tsybakov noise condition. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

Appendix

The appendix is organized as follows: general lemmas regarding geometric properties of the sphere are stated and proved in Section A, lemmas necessary for showing both lower bounds are stated and proved in Section B, the adaptive and random sampling lower bounds are given in Sections C and D, and the random sampling and adaptive upper bounds are given in Sections E and F.

A. General sphere lemmas

A.1. Sine integral bounds

Lemma A.1. *For integral $n \geq 2$,*

$$\frac{7}{4\sqrt{n+1}} \leq \int_0^\pi \sin^n(\theta)d\theta \leq \frac{5}{\sqrt{n+1}} \tag{24}$$

Proof. We first note that, by a standard recursive integration by parts,

$$\int_0^\pi \sin^{2k}(\theta)d\theta = \pi \prod_{i=1}^{k} \frac{2i-1}{2i} \tag{25}$$

$$\tag{26}$$

For both the upper and lower bounds, we first prove the result for even n, then use montonicity to derive a bound for odd n.

Suppose n is even and let $n = 2k$.

$$\int_0^\pi \sin^{2k}(\theta)d\theta = \pi \prod_{i=1}^{k} \frac{2i-1}{2i} \tag{27}$$

$$= \pi \prod_{i=1}^{k} \left(1 - \frac{1}{2i}\right) \tag{28}$$

$$\ln\left(\int_0^\pi \sin^{2k}(\theta)d\theta\right) = \ln(\pi) + \sum_{i=1}^{k} \ln\left(1 - \frac{1}{2i}\right) \tag{29}$$

$$\leq \ln(\pi) + \sum_{i=1}^{k} -\frac{1}{2i} \tag{30}$$

$$= \ln(\pi) - \frac{1}{2}\sum_{i=1}^{k} \frac{1}{i} \tag{31}$$

$$\leq \ln(\pi) - \frac{1}{2}\ln(k+1) \tag{32}$$

$$\leq \ln\left(\frac{\pi}{\sqrt{k+1}}\right) \tag{33}$$

$$\int_0^\pi \sin^{2k}(\theta)d\theta \leq \frac{\pi}{\sqrt{k+1}} = \frac{\pi\sqrt{2}}{\sqrt{n+2}} \leq \frac{5}{\sqrt{n+2}} \tag{34}$$

For odd n, let $n = 2k+1$,

$$\int_0^\pi \sin^{2k+1}(\theta)d\theta \leq \int_0^\pi \sin^{2k}(\theta)d\theta \tag{35}$$

$$\leq \frac{\pi}{\sqrt{k+1}} = \frac{\pi\sqrt{2}}{\sqrt{n+1}} \leq \frac{5}{\sqrt{n+1}} \tag{36}$$

So in either case,

$$\int_0^\pi \sin^{n}(\theta)d\theta \leq \frac{\pi\sqrt{2}}{\sqrt{n+1}} \leq \frac{5}{\sqrt{n+1}} \tag{37}$$

Now for lower bounds, and for even n, let $n = 2k$,

$$\int_0^\pi \sin^{2k}(\theta)d\theta = \pi \prod_{i=1}^{k} \frac{2i-1}{2i} \tag{38}$$

$$= \pi \prod_{i=1}^{k} 1 - \frac{1}{2i} \tag{39}$$

$$\ln\left(\int_0^\pi \sin^{2k}(\theta)d\theta\right) = \ln(\pi) + \sum_{i=1}^{k} \ln(1 - \frac{1}{2i}) \tag{40}$$

$$\geq \ln(\pi) + \sum_{i=1}^{k} (-\frac{1}{2i}) - (-\frac{1}{2i})^2 \tag{41}$$

$$\geq \ln(\pi) - \frac{1}{2}\sum_{i=1}^{k} \frac{1}{i} - \frac{1}{4}\sum_{i=1}^{k} \frac{1}{i^2} \tag{42}$$

$$\geq \ln(\pi) - \frac{1}{2}(1 + \ln(k)) - \frac{1}{4}\frac{\pi^2}{6} \tag{43}$$

$$\geq \ln(\pi) - \frac{1}{2}\ln(k) - \frac{1}{2} - \frac{\pi^2}{24} \tag{44}$$

$$\int_0^\pi \sin^{2k}(\theta)d\theta \geq \frac{\pi}{\sqrt{k}} \exp\left(-\frac{1}{2} - \frac{\pi^2}{24}\right) \tag{45}$$

$$\geq \frac{\pi}{\sqrt{k}} \exp\left(-\frac{1}{2} - \frac{\pi^2}{24}\right) = \frac{\pi\sqrt{2}}{\sqrt{n}} \exp\left(-\frac{1}{2} - \frac{\pi^2}{24}\right) \geq \frac{7}{4\sqrt{n}} \tag{46}$$

For odd n, let $n = 2k - 1$,

$$\int_0^\pi \sin^{2k-1}(\theta)d\theta \geq \int_0^\pi \sin^{2k}(\theta)d\theta \tag{47}$$

$$\geq \frac{\pi}{\sqrt{k}} \exp\left(-\frac{1}{2} - \frac{\pi^2}{24}\right) = \frac{\pi\sqrt{2}}{\sqrt{n+1}} \exp\left(-\frac{1}{2} - \frac{\pi^2}{24}\right) \geq \frac{7}{4\sqrt{n+1}} \tag{48}$$

So in either case,

$$\int_0^\pi \sin^{n}(\theta)d\theta \geq \frac{7}{4\sqrt{n+1}} \tag{49}$$

□

A.2. Spherical Coordinates

We now introduce spherical coordinates which will be used to compute spherical integrals.

$$x_1 = r\sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\sin(\theta_{d-1}) \tag{50}$$
$$x_2 = r\sin(\theta_1)\sin(\theta_2)\dots\sin(\theta_{d-2})\cos(\theta_{d-1}) \tag{51}$$
$$x_3 = r\sin(\theta_1)\sin(\theta_2)\dots\cos(\theta_{d-2}) \tag{52}$$
$$\dots \tag{53}$$
$$x_{d-1} = r\sin(\theta_1)\cos(\theta_2) \tag{54}$$
$$x_d = r\cos(\theta_1) \tag{55}$$

where all angles are in $[0,\pi]$ except θ_{d-1} which is in $[0,2\pi]$ and where the Jacobian determinant is $r^{d-1}\sin^{d-2}(\theta_1)\sin^{d-3}(\theta_2)\dots\sin(\theta_{d-2})$.

A.3. Bounds on CDF of absolute value of coordinate drawn from sphere

Lemma A.2. *Let x be drawn from a d-dimensional sphere of radius ρ, where $d \geq 5$. Then, for $0 \leq \alpha \leq \rho$,*

$$\frac{2}{5}\sqrt{d-1}\left(1-\frac{(d-3)\alpha^2}{2\rho^2}\right)\frac{\alpha}{\rho} \leq \Pr(|x_1|\leq\alpha) \leq \frac{8}{7}\sqrt{d-1}\frac{\alpha}{\rho} \tag{56}$$

Proof. Using symmetry and spherical coordinates,

$$\Pr(|x_1|\leq\alpha) = \Pr(|x_d|\leq\alpha) \tag{57}$$
$$= \Pr(|\rho\cos(\theta_1)|\leq\alpha) \tag{58}$$
$$= \mathbb{E}\left[\mathbf{1}\left[|\cos(\theta_1)|\leq\frac{\alpha}{\rho}\right]\right] \tag{59}$$

Since we are drawing uniformly from the sphere:

$$\mathbb{E}\left[\mathbf{1}\left[|\cos(\theta_1)|\leq\frac{\alpha}{\rho}\right]\right] \tag{60}$$
$$= \frac{\int_{\|x\|=\rho}\mathbf{1}\left[|\cos(\theta_1)|\leq\frac{\alpha}{\rho}\right]dx}{\int_{\|x\|=\rho}dx} \tag{61}$$
$$= \frac{\int_0^\pi\cdots\int_0^\pi\int_0^{2\pi}\mathbf{1}\left[|\cos(\theta_1)|\leq\frac{\alpha}{\rho}\right]\rho^{d-1}\sin^{d-2}(\theta_1)\dots\sin(\theta_{d-2})d\theta_{d-1}\dots d\theta_1}{\int_0^\pi\cdots\int_0^\pi\int_0^{2\pi}\rho^{d-1}\sin^{d-2}(\theta_1)\dots\sin(\theta_{d-2})d\theta_{d-1}\dots d\theta_1} \tag{62}$$
$$= \frac{\rho^{d-1}\int_0^\pi\mathbf{1}\left[|\cos(\theta)\leq\frac{\alpha}{\rho}\right]\sin^{d-2}(\theta)d\theta\int_0^\pi\sin^{d-3}(\theta)d\theta\cdots\int_0^\pi\sin(\theta)d\theta\int_0^{2\pi}d\theta}{\rho^{d-1}\int_0^\pi\sin^{d-2}(\theta)d\theta\int_0^\pi\sin^{d-3}(\theta)d\theta\cdots\int_0^\pi\sin(\theta)d\theta\int_0^{2\pi}d\theta} \tag{63}$$
$$= \frac{\int_0^\pi\mathbf{1}\left[|\cos(\theta)|\leq\frac{\alpha}{\rho}\right]\sin^{d-2}(\theta)d\theta}{\int_0^\pi\sin^{d-2}(\theta)d\theta} \tag{64}$$

From Lemma A.1, we know that the denominator is bounded as

$$\frac{7}{4\sqrt{d-1}} \leq \int_0^\pi \sin^{d-2}(\theta)d\theta \leq \frac{5}{\sqrt{d-1}} \tag{65}$$

To bound the numerator, we use a change of variables of $u = \cos(\theta)$:

$$\int_0^\pi \mathbf{1}\left[|\cos(\theta)| \leq \frac{\alpha}{\rho}\right] \sin^{d-2}(\theta)d\theta = \int_{-1}^1 \mathbf{1}\left[|u| \leq \frac{\alpha}{\rho}\right](1-u^2)^{(d-3)/2}du \tag{66}$$

$$= \int_{-\alpha/\rho}^{\alpha/\rho} (1-u^2)^{(d-3)/2}du \tag{67}$$

To bound the numerator above, we note that the integrand is less than 1:

$$\int_{-\alpha/\rho}^{\alpha/\rho} (1-u^2)^{(d-3)/2}du \leq 2\frac{\alpha}{\rho} \tag{68}$$

To bound the numerator below, we note that $(1-a)^b \geq 1-ab$ for $a \in [0,1]$ and $b \geq 1$ (raise both sides to $\frac{1}{ab}$ then note $(1-x)^{1/x}$ is monotonically decreasing). Then, because $d \geq 5$, $(d-3)/2 \geq 1$ and we have

$$\int_{-\alpha/\rho}^{\alpha/\rho} (1-u^2)^{(d-3)/2}du \geq \int_{-\alpha/\rho}^{\alpha/\rho} (1-u^2(d-3)/2) \tag{69}$$

$$\geq \int_{-\alpha/\rho}^{\alpha/\rho} (1-(\alpha/\rho)^2(d-3)/2) \tag{70}$$

$$= 2\frac{\alpha}{\rho}\left(1 - \frac{(d-3)\alpha^2}{2\rho^2}\right) \tag{71}$$

Putting together the bound on the denominator and the numerator, we arrive at the result. □

A.4. Lower bound for Bayes error

Lemma 4.1. *Suppose $d \geq 5$,*

If $\frac{\sqrt{d-1}}{Mr} \leq 1$, then,

$$err^* \geq \frac{1}{20}\frac{\sqrt{d-1}}{Mr} \tag{72}$$

Proof. Note that

$$\text{err}^* = \text{err}(w^*) \tag{73}$$

$$= \mathbb{E}_{\|x\|=r}[\sigma(w^* \cdot x)\mathbf{1}[w^* \cdot x < 0] + \sigma(-w^* \cdot x)\mathbf{1}[w^* \cdot x > 0]] \tag{74}$$

$$= \mathbb{E}_{\|x\|=r}[\sigma(-|w^* \cdot x|)] \tag{75}$$

Without loss of generality, assume $w^* = Me_1$.

$$\mathbb{E}_{\|x\|=r}[\sigma(-|w^* \cdot x|)] = \mathbb{E}_{\|x\|=r}[\sigma(-|Mx_1|)] \tag{76}$$
$$= \mathbb{E}_{\|x\|=Mr}[\sigma(-|x_1|)] \tag{77}$$
$$\geq \mathbb{E}_{\|x\|=Mr}[\frac{1}{4}\mathbf{1}[|x_1| \leq 1]] \tag{78}$$
$$= \frac{1}{4} \Pr_{\|x\|=Mr}[|x_1| \leq 1] \tag{79}$$
$$\tag{80}$$

Using Lemma A.2 and using the assumption $\frac{\sqrt{d-1}}{Mr} \leq 1$,

$$\frac{1}{4} \Pr_{\|x\|=Mr}[|x_1| \leq 1] \geq \frac{1}{4}\frac{2}{5}\sqrt{d-1}\left(1 - \frac{(d-3)}{2M^2r^2}\right)\frac{1}{Mr} \tag{81}$$
$$\geq \frac{1}{4}\frac{2}{5}\sqrt{d-1}\frac{1}{2}\frac{1}{Mr} \tag{82}$$
$$\geq \frac{1}{20}\frac{\sqrt{d-1}}{Mr} \tag{83}$$

□

A.5. Expression for error

We now derive the error of parameters w given that the true parameters are w^*. Without loss of generality, let $w^* = Me_1$.

$$\text{err}(w) = \int_{\|x\|=r} p(x)[P(y=1|x)\mathbf{1}[x \cdot w < 0] + P_D(y=0|x)\mathbf{1}[x \cdot w > 0]]dx \tag{84}$$
$$= \frac{1}{\int_{\|x\|=r} dx} \int_{\|x\|=r} [\sigma(w^* \cdot x)\mathbf{1}[x \cdot w < 0] + \sigma(-x \cdot w^*)\mathbf{1}[x \cdot w > 0]]dx \tag{85}$$
$$= \frac{1}{\int_{\|x\|=r} dx} \int_{\|x\|=r} 2\sigma(w^* \cdot x)\mathbf{1}[x \cdot w < 0]dx \tag{86}$$
$$= \frac{2}{\int_{\|x\|=r} dx} \int_{x \cdot w < 0, \|x\|=r} \sigma(w^* \cdot x)dx \tag{87}$$

This derivation will be used in the next two lemmas.

A.6. Upper bound for Bayes error

Lemma 4.2. *Suppose* $d \geq 5$.

$$err^* \leq \frac{8}{7}\frac{\sqrt{d-1}}{Mr} \tag{88}$$

Proof. By definition,

$$\text{err}^* = \text{err}(w^*) \tag{89}$$

From equation 87 and assuming (without loss of generality) $w^* = -Me_d$.

$$\text{err}(w^*) = \frac{2}{\int_{\|x\|=r} dx} \int_{x \cdot w^* < 0, \|x\|=r} \sigma(w^* \cdot x) dx \tag{90}$$

$$= \frac{2}{\int_{\|x\|=r} dx} \int_{x_d > 0, \|x\|=r} \sigma(-Mx_d) dx \tag{91}$$

Changing the denominator integral to spherical coordinates:

$$\int_{\|x\|=r} dx = \int_0^\pi \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} r^{d-1} \sin^{d-2}(\theta_1) \sin^{d-3}(\theta_2) \ldots \sin(\theta_{d-2}) d\theta_{d-1} \ldots d_{\theta_1} \tag{92}$$

$$= r^{d-1} \int_0^\pi \sin^{d-2}(\theta) d\theta \int_0^\pi \sin^{d-3}(\theta) d\theta \cdots \int_0^\pi \sin(\theta) d\theta \int_0^{2\pi} d\theta \tag{93}$$

Changing the numerator integral to spherical coordinates:

$$\int_{x_d > 0, \|x\|=r} \sigma(-Mx_d) dx \tag{94}$$

$$= \int_0^{\pi/2} \int_0^\pi \cdots \int_0^\pi \int_0^{2\pi} \exp(-M \cdot r\cos(\theta_1)) r^{d-1} \sin^{d-2}(\theta_1) \sin^{d-3}(\theta_2) \ldots \sin(\theta_{d-2}) d\theta_{d-1} \ldots d\theta_1 \tag{95}$$

$$= r^{d-1} \int_0^{\pi/2} \sigma(-Mr\cos(\theta)) \sin^{d-2}(\theta) d\theta \int_0^\pi \sin^{d-3}(\theta) d\theta \cdots \int_0^\pi \sin(\theta) d\theta \int_0^{2\pi} d\theta \tag{96}$$

Taking the ratio and using a change of variables ($u = \cos(\theta)$), we find,

$$\text{err}(w^*) = \frac{2}{\int_0^\pi \sin^{d-2}(\theta) d\theta} \int_0^{\pi/2} \sigma(-Mr\cos(\theta)) \sin^{d-2}(\theta) d\theta \tag{97}$$

$$= \frac{2}{\int_0^\pi \sin^{d-2}(\theta) d\theta} \int_0^1 \sigma(-Mru)(1-u^2)^{(d-3)/2} du \tag{98}$$

Note that $\sigma(v) \leq \exp(-v)$ and $(1-u^2)^{(d-3)/2} \leq 1$.

So

$$\text{err}(w^*) \leq \frac{2}{\frac{7}{4\sqrt{d-1}}} \int_0^1 \exp(-Mru) du \tag{99}$$

$$= \frac{8\sqrt{d-1}}{7} \frac{1}{Mr} (1 - \exp(-Mr)) \tag{100}$$

$$\leq \frac{8}{7} \frac{\sqrt{d-1}}{Mr} \tag{101}$$

□

A.7. Bounding excess error in terms of angle

Lemma 4.3. *Suppose* $d \geq 5$.

For any w *such that* $\angle(w, w^*) \leq \frac{\pi}{2}$,

$$err(w) - err(w^*) \leq \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\angle(w, w^*)^2 \tag{102}$$

If $Mr\angle(w, w^*) \leq 1$ *and* $\angle(w, w^*) \leq \frac{\pi}{2}$

$$err(w) - err(w^*) \geq \frac{1}{25}\frac{Mr}{\sqrt{d-1}}\angle(w, w^*)^2 \tag{103}$$

Proof. With the expression for error from equation 87, we can find the excess error:

$$\begin{aligned}
&\text{err}(w) - \text{err}(w^*) = && (104)\\
&= \frac{2}{\int_{\|x\|=r} dx}\left(\int_{x\cdot w<0,\|x\|=r} \sigma(w^*\cdot x)dx - \int_{x\cdot w^*<0,\|x\|=r} \sigma(w^*\cdot x)dx\right) && (105)\\
&= \frac{2}{\int_{\|x\|=r} dx}\left(\int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} \sigma(w^*\cdot x)dx - \int_{x\cdot w>0,x\cdot w^*<0,\|x\|=r} \sigma(w^*\cdot x)dx\right) && (106)\\
&= \frac{2}{\int_{\|x\|=r} dx}\left(\int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} \sigma(w^*\cdot x)dx - \int_{-x\cdot w>0,-x\cdot w^*<0,\|x\|=r} \sigma(w^*\cdot (-x))dx\right) && (107)\\
&= \frac{2}{\int_{\|x\|=r} dx}\left(\int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} \sigma(w^*\cdot x)dx - \int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} (1-\sigma(w^*\cdot x))dx\right) && (108)\\
&= \frac{2}{\int_{\|x\|=r} dx}\int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} (2\sigma(w^*\cdot x)-1)dx && (109)
\end{aligned}$$

Our next strategy is to upper and lower bound the integrand with a linear approximation. This strategy is motivated by the following lemma:

Lemma A.3.

$$\frac{2}{\int_{\|x\|=r} dx}\int_{x\cdot w<0,x\cdot w^*>0,\|x\|=r} c(w^*\cdot x)dx = \frac{cMr}{2\pi}(1-\cos(\angle(w, w^*)))\int_0^\pi \sin^{d-1}(\theta)d\theta \tag{110}$$

Proof. Let us use spherical coordinates and without loss of generality we assume $w^* = Me_1$ and w is in the plane spanned by e_1 and e_2. Without loss of generality, since the misclassification error does not depend on the norm of w, assume w is a unit vector, so $w = \cos(\angle(w, w^*))e_1 + \sin(\angle(w, w^*))e_2$.

Note that the integral is over the set of points x on the r-radius sphere where

$$\begin{aligned}
x\cdot w^* > 0 \wedge x\cdot w < 0 && (111)\\
x_1 > 0 \wedge \cos(\angle(w, w^*))x_1 + \sin(\angle(w, w^*))x_2 < 0 && (112)\\
\sin(\theta_{d-1}) > 0 \wedge \cos(\angle(w, w^*))\sin(\theta_{d-1}) + \sin(\angle(w, w^*))\cos(\theta_{d-1}) < 0 && (113)\\
\sin(\theta_{d-1}) > 0 \wedge \sin(\theta_{d-1} - \angle(w, w^*)) < 0 && (114)\\
\theta_{d-1} \in [0, \angle(w, w^*)] && (115)
\end{aligned}$$

Thus,

$$\frac{2}{\int_{\|x\|=r} dx} \int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r} c(w^*\cdot x)dx = \tag{116}$$

$$= 2\frac{\int_0^\pi \int_0^\pi \cdots \int_0^{\angle(w,w_D^*)} c(w^*\cdot x) r^{d-1}\sin^{d-2}(\theta_1)\sin^{d-3}(\theta_2)\ldots\sin(\theta_{d-2})d\theta_{d-1}\ldots d\theta_1}{\int_0^\pi \int_0^\pi \cdots \int_0^{2\pi} r^{d-1}\sin^{d-2}(\theta_1)\sin^{d-3}(\theta_2)\ldots\sin(\theta_{d-2})d\theta_{d-1}\ldots d\theta_1} \tag{117}$$

$$\tag{118}$$

Note that $w^*\cdot x = Mx_1 = Mr\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-1})$

$$\frac{2}{\int_{\|x\|=r} dx} \int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r} c(w^*\cdot x)dx = \tag{119}$$

$$= 2c\frac{\int_0^\pi \int_0^\pi \cdots \int_0^{\angle(w,w^*)} Mr^d\sin^{d-1}(\theta_1)\sin^{d-2}(\theta_2)\ldots\sin^2(\theta_{d-2})\sin(\theta_{d-1})d\theta_{d-1}\ldots d\theta_1}{\int_0^\pi \int_0^\pi \cdots \int_0^{2\pi} r^{d-1}\sin^{d-2}(\theta_1)\sin^{d-3}(\theta_2)\ldots\sin(\theta_{d-2})d\theta_{d-1}\ldots d\theta_1} \tag{120}$$

$$= 2cMr\frac{\int_0^\pi \sin^{d-1}(\theta)d\theta\cdots\int_0^\pi \sin^2(\theta)d\theta\int_0^{\angle(w,w^*)}\sin(\theta)d\theta}{\int_0^\pi \sin^{d-2}(\theta)d\theta\cdots\int_0^\pi \sin(\theta)d\theta\int_0^{2\pi} d\theta} \tag{121}$$

$$= 2cMr\frac{\int_0^\pi \sin^{d-1}(\theta)d\theta(1-\cos(\angle(w,w^*)))}{2\cdot 2\pi} \tag{122}$$

$$= \frac{cMr}{2\pi}(1-\cos(\angle(w,w^*)))\int_0^\pi \sin^{d-1}(\theta)d\theta \tag{123}$$

□

We now return to the proof of Lemma 4.3.

Bounding it above:

Note that $2\sigma(u)-1 \le u/2$ (for $u\ge 0$). Note that since $\angle(w,w^*)\le\frac{\pi}{2}$, $w^*\cdot x\ge 0$. Thus,

$$\text{err}(w)-\text{err}(w^*) = \frac{2}{\int_{\|x\|=r} dx}\int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r}(2\sigma(w^*\cdot x)-1)dx \tag{124}$$

$$\le \frac{2}{\int_{\|x\|=r} dx}\int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r}\frac{1}{2}(w^*\cdot x)dx \tag{125}$$

$$= \frac{1}{2}\frac{Mr}{2\pi}(1-\cos(\angle(w,w^*)))\int_0^\pi \sin^{d-1}(\theta)d\theta \tag{126}$$

$$\le \frac{1}{2}\frac{Mr}{2\pi}\frac{\angle(w,w^*)^2}{2}\frac{5}{\sqrt{d}} \tag{127}$$

$$\le \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\angle(w,w^*)^2 \tag{128}$$

Bounding it below:

Note that $2\sigma(u)-1\ge 5u/11$ for $u\in[0,1]$.

By assumption, $Mr\angle(w,w^*)\le 1$. Then, $w^*\cdot x = Mr\sin(\theta_1)\sin(\theta_2)\ldots\sin(\theta_{d-1})\le Mr\sin(\theta_{d-1})\le Mr\angle(w,w^*)\le 1$. Further, since $\angle(w,w^*)\le\frac{\pi}{2}$, $w^*\cdot x\ge 0$. So $w^*\cdot x\in[0,1]$.

$$\text{err}(w) - \text{err}(w^*) = \frac{2}{\int_{\|x\|=r} dx} \int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r} (2\sigma(w^*\cdot x) - 1)dx \tag{129}$$

$$\geq \frac{2}{\int_{\|x\|=r} dx} \int_{x\cdot w<0, x\cdot w^*>0, \|x\|=r} \frac{5}{11}(w^*\cdot x)dx \tag{130}$$

$$= \frac{5}{11}\frac{Mr}{2\pi}(1-\cos(\angle(w,w^*)))\int_0^\pi \sin^{d-1}(\theta)d\theta \tag{131}$$

$$\geq \frac{5}{11}\frac{Mr}{2\pi}\frac{\angle(w,w^*)^2}{3}\frac{7}{4\sqrt{d}} \tag{132}$$

$$\geq \frac{1}{25}\frac{Mr}{\sqrt{d-1}}\angle(w,w^*)^2 \tag{133}$$

□

B. Lower bound setup and lemmas

Let $\mathcal{F}$ be the set of (measurable) decision rules $f : \mathcal{X} \to \mathcal{Y} = \{0, 1\}$. Note we use $\mathcal{Y} = \{0, 1\}$ here for notational simplicity, but the same results hold for $\mathcal{Y} = \{-1, 1\}$.

Suppose we have a finite set of conditional distributions Π, where $\pi : \mathcal{X} \to [0, 1]$.

We can define a dissimilarity between decision rules $f \in \mathcal{F}$ and distributions $\pi \in \Pi$ as following:

$$\rho(f, \pi) = \mathbb{E}[f(x)(1 - \pi(x)) + (1 - f(x))\pi(x)] \tag{134}$$

Intuitively, $\rho(f, \pi)$ is the zero-one loss or misclassification error if we predict $f(x)$ and the probability of $y = 1$ is $\pi(x)$.

We study a setup where there is a true label distribution π^* (which generates data) and an estimated classifier $\hat{f}$ (computed from data).

Motivated by this observation, let $\text{err}(f) = \rho(f, \pi^*)$, and $\text{err}^* = \inf_{f\in\mathcal{F}} \rho(f, \pi^*)$.

Define the closest distribution to $\hat{f}$ as $\hat{\pi}$.

$$\hat{\pi} = \operatorname*{argmin}_{\pi\in\Pi} \rho(\hat{f}, \pi) \tag{135}$$

We proceed with a general lemma that the error of a hypothesis on a mixture of distributions is the mixture of the error on the distributions.

Lemma B.1. *If* $\sum_i \alpha_i = 1$*, then*

$$\rho\left(f, \sum_i \alpha_i \pi_i\right) = \sum_i \alpha_i \rho(f, \pi_i) \tag{136}$$

Proof.

$$\rho\left(f, \sum_i \alpha_i \pi_i\right) = \mathbb{E}\left[f(x)\left(1 - \sum_i \alpha_i \pi_i(x)\right) + (1 - f(x))\left(\sum_i \alpha_i \pi_i(x)\right)\right] \tag{137}$$

$$= \mathbb{E}\left[f(x)\left(\sum_i \alpha_i (1 - \pi_i(x))\right) + (1 - f(x))\left(\sum_i \alpha_i \pi_i(x)\right)\right] \tag{138}$$

$$= \sum_i \alpha_i \mathbb{E}[f(x)(1 - \pi_i(x)) + (1 - f(x))\pi_i(x)] \tag{139}$$

$$= \sum_i \alpha_i \rho(f, \pi_i) \tag{140}$$

□

B.1. Excess error difference

Define π_w be the conditional label distribution corresponding to logistic regression with weights w: $\pi_w(x) = \Pr(y = 1|x) = \sigma(w \cdot x)$.

Lemma B.2. *Let $d \geq 5$. Suppose $w_1, w_2 \in \mathbb{R}^d$ such that $\|w_1\| = \|w_2\| = M$, $Mr\angle(w_1, w_2) \leq 2$, and $\angle(w_1, w_2) < \pi$. Let x be drawn uniformly from a radius r sphere in d dimensions. If $\pi_1 = \pi_{w_1}$ and $\pi_2 = \pi_{w_2}$,*

$$\inf_f \rho\left(f, \frac{\pi_1 + \pi_2}{2}\right) - \frac{\inf_f \rho(f, \pi_1) + \inf_f \rho(f, \pi_2)}{2} > \frac{1}{101} \frac{Mr}{\sqrt{d-1}} \angle(w_1, w_2)^2 \tag{141}$$

Proof. Define the decision rule $f_w(x) = \mathbf{1}[w \cdot x \geq 0]$. Note that the infima $\inf_f \rho(f, \pi_1)$ and $\inf_f \rho(f, \pi_2)$ are attained at $f_1 = f_{w_1}$ and $f_2 = f_{w_2}$.

Define $\overline{\pi} = \frac{\pi_1 + \pi_2}{2}$.

For $\overline{\pi}$, the infimum is attained at $\overline{f}$:

$$\overline{f}(x) = \mathbf{1}\left[\overline{\pi}(x) \geq \frac{1}{2}\right] \tag{142}$$

$$= \mathbf{1}\left[\frac{\sigma(w_1 \cdot x) + \sigma(w_2 \cdot x)}{2} \geq \frac{1}{2}\right] \tag{143}$$

$$= \mathbf{1}\left[\sigma(w_1 \cdot x) \geq 1 - \sigma(w_2 \cdot x)\right] \tag{144}$$

$$= \mathbf{1}[\sigma(w_1 \cdot x) \geq \sigma(-w_2 \cdot x)] \tag{145}$$

$$= \mathbf{1}[w_1 \cdot x \geq -w_2 \cdot x] \tag{146}$$

$$= \mathbf{1}\left[(w_1 + w_2) \cdot x \geq 0\right] \tag{147}$$

Let $\overline{w} = w_1 + w_2$, so $\overline{f} = f_{\overline{w}}$.

Thus,

$$\inf_f \rho(f, \overline{\pi}) - \frac{\inf_f \rho(f, \pi_1) + \inf_f \rho(f, \pi_2)}{2} = \rho(\overline{f}, \overline{\pi}) - \frac{\rho(f_1, \pi_1) + \rho(f_2, \pi_2)}{2} \tag{148}$$

$$= \frac{\rho(\overline{f}, \pi_1) + \rho(\overline{f}, \pi_2)}{2} - \frac{\rho(f_1, \pi_1) + \rho(f_2, \pi_2)}{2} \tag{149}$$

$$= \frac{1}{2}\left(\left[\rho(\overline{f}, \pi_1) - \rho(f_1, \pi_1)\right] + \left[\rho(\overline{f}, \pi_2) - \rho(f_2, \pi_2)\right]\right) \tag{150}$$

Note that the terms in the square brackets are the excess error of $\overline{f}$ if π_i were the true distribution. Furthermore, π_i is a logistic conditional label distribution and $\overline{f}$ is a linear classifier, so we can use Lemma 4.3.

Note that by symmetry, $\angle(\overline{w}, w_1) = \angle(\overline{w}, w_2)$.

Furthermore,

$$\cos(2\angle(\overline{w}, w_1)) = 2\cos(\angle(\overline{w}, w_1))^2 - 1 \tag{151}$$
$$= 2\left(\frac{\overline{w} \cdot w_1}{\|\overline{w}\|\|w_1\|}\right)^2 - 1 \tag{152}$$
$$= 2\frac{(\|w_1\|^2 + w_1 \cdot w_2)^2}{(\|w_1\|^2 + 2w_1 \cdot w_2 + \|w_2\|^2)\|w_1\|^2} - 1 \tag{153}$$
$$= \frac{(M^2 + w_1 \cdot w_2)^2}{(M^2 + w_1 \cdot w_2)M^2} - 1 \tag{154}$$
$$= \frac{M^2 + w_1 \cdot w_2}{M^2} - 1 \tag{155}$$
$$= \frac{w_1 \cdot w_2}{M^2} \tag{156}$$
$$= \cos(\angle(w_1, w_2)) \tag{157}$$

Thus, $2\angle(\overline{w}, w_1) = \angle(w_1, w_2)$. Since $Mr\angle(w_1, w_2) \leq 2$, $Mr\angle(\overline{w}, w_1) = Mr\angle(\overline{w}, w_2) \leq 1$ and since $\angle(w_1, w_2) < \pi$, $\angle(\overline{w}, w_1) = \angle(\overline{w}, w_1) < \frac{\pi}{2}$. Thus, we meet the conditions of Lemma 4.3.

Using Lemma 4.3,

$$\frac{1}{2}\left(\left[\rho(\overline{f}, \pi_1) - \rho(f_1, \pi_1)\right] + \left[\rho(\overline{f}, \pi_2) - \rho(f_2, \pi_2)\right]\right) \geq \frac{1}{2}\left(\frac{1}{25}\frac{Mr}{\sqrt{d-1}}\angle(\overline{w}, w_1)^2 + \frac{1}{25}\frac{Mr}{\sqrt{d-1}}\angle(\overline{w}, w_2)^2\right) \tag{158}$$
$$= \frac{1}{100}\frac{Mr}{\sqrt{d-1}}\angle(w_1, w_2)^2 \tag{159}$$
$$> \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\angle(w_1, w_2)^2 \tag{160}$$

□

B.2. Excess error and distribution similarity

Recall Π is a set of conditional label distributions. Let $\mathcal{S} \subset \Pi \times \Pi$ be a set of "similar" pairs of distributions.

Lemma B.3. *Suppose* $\inf_f \rho(f, \pi) = \inf_f \rho(f, \pi')$ *for all* $\pi, \pi' \in \Pi$. *Fix* $a > 0$.

If, for all $\pi_1, \pi_2 \in \Pi$ *where* $(\pi_1, \pi_2) \notin \mathcal{S}$,

$$\inf_f \rho\left(f, \frac{\pi_1 + \pi_2}{2}\right) - \frac{\inf_f \rho(f, \pi_1) + \inf_f \rho(f, \pi_2)}{2} > a \tag{161}$$

then

$$err(\hat{f}) - err^* \leq a \implies (\hat{\pi}, \pi^*) \in \mathcal{S} \tag{162}$$

Proof. For any $\hat{f}$ and for any π_1, π_2 such that $(\pi_1, \pi_2) \notin \mathcal{S}$,

$$\rho\left(\hat{f}, \frac{\pi_1 + \pi_2}{2}\right) - \frac{\inf_f \rho(f, \pi_1) + \inf_f \rho(f, \pi_2)}{2} > a \tag{163}$$

$$\frac{\rho(\hat{f}, \pi_1) + \rho(\hat{f}, \pi_2)}{2} - \frac{\inf_f \rho(f, \pi_1) + \inf_f \rho(f, \pi_2)}{2} > a \tag{164}$$

$$\left(\rho(\hat{f}, \pi_1) - \inf_f \rho(f, \pi_1)\right) + \left(\rho(\hat{f}, \pi_2) - \inf_f \rho(f, \pi_2)\right) > 2a \tag{165}$$

Thus, if $\rho(\hat{f}, \pi_1) - \inf_f \rho(f, \pi_1) \leq a$, it must be the case that $\rho(\hat{f}, \pi_2) - \inf_f \rho(f, \pi_2) > a$.

Therefore,

$$\text{err}(\hat{f}) - \text{err}^* \leq a \implies \rho(\hat{f}, \pi^*) - \inf_f \rho(f, \pi^*) \leq a \tag{166}$$

$$\implies \forall \pi' : (\pi^*, \pi') \notin \mathcal{S}, \rho(\hat{f}, \pi') - \inf_f \rho(f, \pi') > a \tag{167}$$

$$\implies \forall \pi' : (\pi^*, \pi') \notin \mathcal{S}, \rho(\hat{f}, \pi^*) < \rho(\hat{f}, \pi') \tag{168}$$

$$\implies \forall \pi' : (\pi^*, \pi') \notin \mathcal{S}, \hat{\pi} \neq \pi' \tag{169}$$

$$\implies (\pi^*, \hat{\pi}) \in \mathcal{S} \tag{170}$$

Thus, if the "excess error" of $\hat{f}$ is low enough, then the estimated distribution $\hat{\pi}$ is similar to π^*.

$$\text{err}(\hat{f}) - \text{err}^* \leq a \implies (\hat{\pi}, \pi^*) \in \mathcal{S} \tag{171}$$

□

B.3. Fano's inequality

Suppose we have a finite set of objects $\mathcal{V}$. Suppose we have a random variable $V^* \in \mathcal{V}$ uniformly at random drawn from $\mathcal{V}$ that we attempt to estimate by a random variable $\hat{V}$.

We are interested in upper bounding the probability that $\hat{V}$ and V^* are "similar". In other words, let $\mathcal{S} \subset \mathcal{V} \times \mathcal{V}$. Then we wish to upper bound the probability that $(V^*, \hat{V}) \in \mathcal{S}$.

Define,

$$N_{\max}(\mathcal{S}) = \max_{\hat{v} \in \mathcal{V}} \sum_{v^* \in \mathcal{V}} \mathbf{1}[(v^*, \hat{v}) \in \mathcal{S}] \tag{172}$$

Then,

Lemma B.4 (Theorem 2 from Scarlett & Cevher (2019))**.**

$$\Pr((V^*, \hat{V}) \in \mathcal{S}) \leq \frac{I(\hat{V}; V^*) + \ln(2)}{\ln \frac{|\mathcal{V}|}{N_{max}(\mathcal{S})}} \tag{173}$$

This lemma is proved in Duchi & Wainwright (2013).

B.4. Instantiation of Π and $\mathcal{S}$

Given the previous lemmas, we are ready to instantiate $\Pi = \mathcal{V}$ and $\mathcal{S}$.

Let $\varepsilon > 0$. Let $\mathcal{W} = \frac{M}{\sqrt{1+(d-1)\varepsilon^2}}(1, \pm\varepsilon, \cdots \pm\varepsilon)$ be a set of $|\mathcal{W}| = 2^{d-1}$ vectors of dimension d. Further note that for any $w \in \mathcal{W}$, $\|w\| = M$. Let $\mathcal{V} = \Pi = \{\pi_w : w \in \mathcal{W}\}$

Now, we prove a small lemma bounding the maximum angle between two vectors in $\mathcal{W}$.

Lemma B.5.

$$\max_{w\in\mathcal{W}} \angle(w, e_1) \leq 2\sqrt{(d-1)\varepsilon^2} \tag{174}$$

and thus, by triangle inequality

$$\max_{w_1,w_2\in\mathcal{W}} \angle(w_1, w_2) \leq 4\sqrt{(d-1)\varepsilon^2} \tag{175}$$

Proof. For any $w \in \mathcal{W}$,

$$\cos(\angle(w_1, e_1)) = \frac{w \cdot e_1}{\|w\|\|e_1\|} \tag{176}$$

$$= \frac{1}{\sqrt{1+(d-1)\varepsilon^2}} \tag{177}$$

$$1 - \frac{\angle(w_1, e_1)^2}{5} \geq \frac{1}{\sqrt{1+(d-1)\varepsilon^2}} \tag{178}$$

$$\angle(w_1, e_1)^2 \leq 5\left(1 - \frac{1}{\sqrt{1+(d-1)\varepsilon^2}}\right) \tag{179}$$

$$\leq \frac{5}{2}(d-1)\varepsilon^2 \tag{180}$$

$$\angle(w_1, e_1) \leq 2\sqrt{(d-1)\varepsilon^2} \tag{181}$$

The second to last line follows from noting that $1 - \frac{1}{\sqrt{1+a}} \leq \frac{a}{2}$ for $a > 0$. □

Let $\mathcal{S} = \left\{(\pi_{w_1}, \pi_{w_2}) : w_1, w_2 \in \mathcal{W}, \angle(w_1, w_2)^2 \leq \frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right\} \subset \Pi \times \Pi$.

Define the Hamming distance between two vectors to be the number of dimensions with different values: $\text{Hamm}(w_1, w_2) = \sum_{i=1}^{d} \mathbf{1}[(w_1)_i \neq (w_2)_i]$. This allows for a lemma connecting the Hamming distance to the angle for vectors in $\mathcal{W}$.

Lemma B.6. *For any $w_1, w_2 \in \mathcal{W}$*

$$\angle(w_1, w_2)^2 \geq \frac{4\textit{Hamm}(w_1, w_2)\varepsilon^2}{1+(d-1)\varepsilon^2} \tag{182}$$

Proof.

$$\cos(\angle(w_1, w_2)) = \frac{w_1 \cdot w_2}{\|w_1\|\|w_2\|} \tag{183}$$

$$= \frac{1}{M^2}\left(\frac{M^2}{1+(d-1)\varepsilon^2}\right)\left(1 + ((d-1) - \text{Hamm}(w_1, w_2))(\varepsilon^2) + (\text{Hamm}(w_1, w_2))(-\varepsilon^2)\right) \tag{184}$$

$$= \left(\frac{1}{1+(d-1)\varepsilon^2}\right)\left(1 + (d-1)\varepsilon^2 - 2\text{Hamm}(w_1, w_2)\varepsilon^2\right) \tag{185}$$

$$= 1 - \frac{2\text{Hamm}(w_1, w_2)\varepsilon^2}{1+(d-1)\varepsilon^2} \tag{186}$$

$$1 - \frac{\angle(w_1, w_2)^2}{2} \leq 1 - \frac{2\text{Hamm}(w_1, w_2)\varepsilon^2}{1+(d-1)\varepsilon^2} \tag{187}$$

$$\angle(w_1, w_2)^2 \geq \frac{4\text{Hamm}(w_1, w_2)\varepsilon^2}{1+(d-1)\varepsilon^2} \tag{188}$$

□

This Hamming distance lemma is important in proving the following Lemma:

Lemma B.7.

$$N_{max}(\mathcal{S}) \leq |\mathcal{W}| \exp\left(-\frac{d-1}{16}\right) \tag{189}$$

Proof.

$$N_{\max}(\mathcal{S}) = \max_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} \mathbf{1}\left[\angle(w_1, w_2)^2 \leq \frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right] \tag{190}$$

$$\leq \max_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} \mathbf{1}\left[\frac{4\text{Hamm}(w_1, w_2)\varepsilon^2}{1+(d-1)\varepsilon^2} \leq \frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right] \tag{191}$$

$$= \max_{w_1 \in \mathcal{W}} \sum_{w_2 \in \mathcal{W}} \mathbf{1}\left[\text{Hamm}(w_1, w_2) \leq \frac{d-1}{4}\right] \tag{192}$$

$$= |\mathcal{W}| \Pr\left(\text{Binomial}\left(d-1, \frac{1}{2}\right) \leq \frac{d-1}{4}\right) \tag{193}$$

$$\leq |\mathcal{W}| \exp\left(-\frac{d-1}{16}\right) \tag{194}$$

Where the last line follows from a Chernoff bound. □

B.5. Putting the parts together

Let $\pi^* = \pi_{w^*}$ be drawn uniformly from Π defined by $\mathcal{W} = \frac{M}{\sqrt{1+(d-1)\varepsilon^2}}(1, \pm\varepsilon, \ldots, \pm\varepsilon)$. Let $\hat{f}$ be a (random) estimated decision rule that possibly depends on π^*.

Lemma B.8. *Let $d \geq 24$. For any estimator $\hat{f}$, if there exists an $\varepsilon > 0$ small enough so that $4(d-1)\varepsilon^2 \leq 1$, $2Mr\sqrt{d-1}\varepsilon \leq 1$, and $I(\hat{f}; \pi^*) \leq \frac{d-1}{64}$, then,*

$$\mathbb{E}[err(\hat{f})] - err^* \geq \frac{1}{808} Mr\sqrt{d-1}\varepsilon^2 \tag{195}$$

Proof. From Lemma B.5,

$$\max_{w_1,w_2\in\mathcal{W}} \angle(w_1,w_2) \le 4\sqrt{d-1}\varepsilon \tag{196}$$

Therefore, $Mr\max_{w_1,w_2\in\mathcal{W}}\angle(w_1,w_2) \le Mr(4\sqrt{d-1}\varepsilon) \le 2$

Furthermore, $\max_{w_1,w_2\in\mathcal{W}}\angle(w_1,w_2) \le 4\sqrt{d-1}\epsilon \le 2$

From Lemma B.2, for $w_1,w_2\in\mathcal{W}$ and since $Mr\angle(w_1,w_2)\le 2$ and $\angle(w_1,w_2)\le 2<\pi$, for $\pi_1=\pi_{w_1}$ and $\pi_2=\pi_{w_2}$,

$$\inf_f \rho\left(f,\frac{\pi_1+\pi_2}{2}\right) - \frac{\inf_f\rho(f,\pi_1)+\inf_f\rho(f,\pi_2)}{2} > \frac{1}{120}\frac{Mr}{\sqrt{d-1}}\angle(w_1,w_2)^2 \tag{197}$$

By the definition of $\mathcal{S}$, if $(\pi_1,\pi_2)\notin\mathcal{S}$, $\angle(w_1,w_2)^2 > \frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}$ and thus,

$$\inf_f \rho\left(f,\frac{\pi_1+\pi_2}{2}\right) - \frac{\inf_f\rho(f,\pi_1)+\inf_f\rho(f,\pi_2)}{2} > \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2} \tag{198}$$

Since all $w\in\mathcal{W}$ have the same norm, $\inf_f\rho(f,\pi_w)$ are all equal. Therefore, by Lemma B.3,

$$\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2} \implies (\hat{\pi},\pi^*)\in\mathcal{S} \tag{199}$$

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right) \le \Pr\left((\hat{\pi},\pi^*)\in\mathcal{S}\right) \tag{200}$$

Then, by Lemma B.4

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right) \le \frac{I(\hat{\pi};\pi^*)+\ln 2}{\ln\frac{|\mathcal{W}|}{N_{\max}(\mathcal{S})}} \tag{201}$$

By Lemma B.7 and noting that by the data processing inequality $I(\hat{\pi};\pi^*)\le I(\hat{f};\pi^*)$,

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{101}\frac{Mr}{\sqrt{d-1}}\frac{(d-1)\varepsilon^2}{1+(d-1)\varepsilon^2}\right) \le \frac{I(\hat{f};\pi^*)+\ln 2}{(d-1)/16} \tag{202}$$

Since $(d-1)\varepsilon^2\le 1$,

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{202}Mr\sqrt{d-1}\varepsilon^2\right) \le \frac{I(\hat{\pi};\pi^*)+\ln 2}{(d-1)/16} \tag{203}$$

Since $d\ge 24$, $\frac{16\ln 2}{d-1}\le\frac{1}{2}$.

By assumption, $I(\hat{f};\pi^*)\le\frac{d-1}{64}$, $\frac{16I(\hat{f},\pi^*)}{d-1}\le\frac{1}{4}$

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \le \frac{1}{202}Mr\sqrt{d-1}\varepsilon^2\right) \le \frac{1}{4}+\frac{1}{2} = \frac{3}{4} \tag{204}$$

Looking at the complementary event,

$$\Pr\left(\text{err}(\hat{f}) - \text{err}^* \geq \frac{1}{202} Mr\sqrt{d-1}\varepsilon^2\right) \geq \frac{1}{4} \tag{205}$$

By a Markov-style inequality,

$$\mathbb{E}[\text{err}(\hat{f})] - \text{err}^* \geq \frac{1}{808} Mr\sqrt{d-1}\varepsilon^2 \tag{206}$$

Finally, since the expectation includes the draw of w^*, the worst-case w^* has at least as much excess error. □

B.6. Mutual information lemmas

We now switch to the objective of bounding the mutual information between labels and the distributions that they are drawn from.

Lemma B.9. *Let P and Q be Bernoulli random variables. Then,*

$$D_{KL}(P\|Q) \leq \frac{(\mathbb{E}[P] - \mathbb{E}[Q])^2}{\mathbb{E}[Q](1-\mathbb{E}[Q])} \tag{207}$$

Proof.

$$D_{KL}(P\|Q) = \Pr(P=1)\ln\frac{\Pr(P=1)}{\Pr(Q=1)} + \Pr(P=0)\ln\frac{\Pr(P=0)}{\Pr(Q=0)} \tag{208}$$

$$= \mathbb{E}[P]\ln\frac{\mathbb{E}[P]}{\mathbb{E}[Q]} + (1-\mathbb{E}[P])\ln\frac{1-\mathbb{E}[P]}{1-\mathbb{E}[Q]} \tag{209}$$

$$\leq \mathbb{E}[P]\left(\frac{\mathbb{E}[P]}{\mathbb{E}[Q]} - 1\right) + (1-\mathbb{E}[P])\left(\frac{1-\mathbb{E}[P]}{1-\mathbb{E}[Q]} - 1\right) \tag{210}$$

$$= \frac{(\mathbb{E}[P] - \mathbb{E}[Q])^2}{\mathbb{E}[Q](1-\mathbb{E}[Q])} \tag{211}$$

□

Consider a parametrized family of conditional probability models $\{\pi_w : w \in \mathcal{W}\}$ on input space $\mathcal{X}$ and binary label space $\mathcal{Y} = \{0, 1\}$.

$$\pi_w(x) = \Pr(Y = 1|x; w) \tag{212}$$

For a fixed $x \in \mathcal{X}$, consider the process:

- $w \sim U(\mathcal{W})$
- $Y \sim \text{Bernoulli}(\pi_w(x))$

Define $I_x(Y; w)$ as the mutual information between Y and w.

Lemma B.10.

$$I_x(Y; w) \leq \frac{\mathit{Var}_w(p_w(x))}{\mathbb{E}_w[p_w(x)(1-p_w(x))]} \tag{213}$$

Proof.

$$I_x(Y;w) = \mathbb{E}_w[D_{KL}(p_{Y|w}\|p_Y)] \tag{214}$$
$$\leq \mathbb{E}_w\left[\frac{(\mathbb{E}[Y|w]-\mathbb{E}[Y])^2}{\mathbb{E}[Y](1-\mathbb{E}[Y])}\right] \tag{215}$$
$$= \frac{\mathbb{E}_w[(\mathbb{E}[Y|w]-\mathbb{E}[Y])^2]}{\mathbb{E}[Y](1-\mathbb{E}[Y])} \tag{216}$$
$$= \frac{\mathbb{E}_w[(p_w(x)-\mathbb{E}_w[p_w(x)])^2]}{\mathbb{E}_w[p_w(x)](1-\mathbb{E}_w[p_w(x)])} \tag{217}$$
$$= \frac{\text{Var}_w(p_w(x))}{\mathbb{E}_w[p_w(x)](1-\mathbb{E}_w[p_w(x)])} \tag{218}$$
$$\leq \frac{\text{Var}_w(p_w(x))}{\mathbb{E}_w[p_w(x)(1-p_w(x))]} \tag{219}$$

□

where the last line follows from Jensen's inequality.

B.6.1. Specializing for logistic regression

Specializing to the case of logistic regression:

Lemma B.11. *Let Σ_w be the covariance of $w \sim U(\mathcal{W})$. Let $C(\mathcal{W})$ be the convex hull of $\mathcal{W}$ and $D(C(\mathcal{W}))$ be the diameter of the convex hull of $\mathcal{W}$. Then, if $\pi_w(x) = \sigma(w \cdot x)$,*

$$I_x(Y;w) \leq 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right]\exp(D(C(\mathcal{W}))\|x\|)x^T\Sigma_w x \tag{220}$$

Proof.

$$I_x(Y;w) \leq \frac{\frac{1}{2}\mathbb{E}_{w,w'}[(\sigma(w\cdot x)-\sigma(w'\cdot x))^2]}{\mathbb{E}_w[\sigma(w\cdot x)(1-\sigma(w\cdot x))]} \tag{221}$$
$$\leq \frac{\frac{1}{2}\mathbb{E}_{w,w'}[(\sigma'(w''\cdot x)x\cdot(w-w'))^2]}{\min_{w\in\mathcal{W}}\sigma(w\cdot x)(1-\sigma(w\cdot x))} \tag{222}$$

where w'' is on the line between w and w'. Define $C(\mathcal{W})$ as the convex hull of $\mathcal{W}$, then note that

$$I_x(Y;w) \leq \frac{\max_{w\in C(\mathcal{W})}\sigma'(w\cdot x)^2\frac{1}{2}\mathbb{E}_{w,w'}[(x\cdot(w-w'))^2]}{\min_{w\in\mathcal{W}}\sigma(w\cdot x)(1-\sigma(w\cdot x))} \tag{223}$$

Define $\psi(u) = \sigma(u)(1-\sigma(u))$, and note that $\sigma'(u) = \psi(u)$.

$$I_x(Y;w) \leq \left[\max_{w\in C(\mathcal{W})}\psi(w\cdot x)\right]\frac{\max_{w\in C(\mathcal{W})}\psi(w\cdot x)}{\min_{w\in\mathcal{W}}\psi(w\cdot x)}x^T\Sigma_w x \tag{224}$$

Next, note that $\psi(u) \leq \frac{1}{4}$ and $\frac{1}{4}\exp(-|u|) \leq \psi(u) \leq \exp(-|u|)$ for all u.

$$I_x(Y;w) \le \left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] \frac{\max_{w\in C(\mathcal{W})} \exp(-|w\cdot x|)}{\min_{w\in\mathcal{W}} \frac{1}{4}\exp(-|w\cdot x|)} x^T\Sigma_w x \tag{225}$$

$$\le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] \max_{w,w'\in C(\mathcal{W})} \exp(|w\cdot x| - |w'\cdot x|) x^T\Sigma_w x \tag{226}$$

$$\le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] \max_{w,w'\in C(\mathcal{W})} \exp(|(w-w')\cdot x|) x^T\Sigma_w x \tag{227}$$

$$\le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] \exp(D(C(\mathcal{W}))\|x\|) x^T\Sigma_w x \tag{228}$$

$$\tag{229}$$

□

Lemma B.12. *Suppose w is drawn uniformly at random from $\mathcal{W} = \frac{M}{\sqrt{1+(d-1)\varepsilon^2}}(1,\pm\varepsilon,\dots,\pm\varepsilon)$ and x is on a radius r sphere ($\|x\| = r$)*

$$I_x(Y;w) \le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] M^2\varepsilon^2 r^2 \exp\left(2\sqrt{d-1}\varepsilon M r\right) \tag{230}$$

Proof. For the particular setting of $\mathcal{W}$,

$$\Sigma_w = \frac{M^2\varepsilon^2}{1+(d-1)\varepsilon^2}(I_d - e_1e_1^T) \tag{231}$$

$$D(C(\mathcal{W})) = \frac{M}{\sqrt{1+(d-1)\varepsilon^2}}(2\sqrt{d-1}\varepsilon) \tag{232}$$

and thus, by Lemma B.11,

$$I_x(Y;w) \le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] \frac{M^2\varepsilon^2}{1+(d-1)\varepsilon^2}\|x\|^2 \exp\left(\frac{M}{\sqrt{1+(d-1)\varepsilon^2}} 2\sqrt{d-1}\varepsilon\|x\|\right) \tag{233}$$

$$\le 4\left[\max_{w\in C(\mathcal{W})} \psi(w\cdot x)\right] M^2 r^2\varepsilon^2 \exp\left(2\sqrt{d-1}Mr\varepsilon\right) \tag{234}$$

□

C. Adaptive lower bound

Let $\pi^* = \pi_{w^*}$ be drawn uniformly from Π defined by $\mathcal{W} = \frac{M}{\sqrt{1+(d-1)\varepsilon^2}}(1,\pm\varepsilon,\dots,\pm\varepsilon)$. Let there be a strategy to collect (possibly adaptively) n data points $\{(X_i,Y_i)\}_{i=1}^n$ where X_i is on the surface of a radius r sphere and Y_i is the associated label generated under π^*. Let $\hat{f}$ be a decision rule based on the data, so that $\hat{f}$ is conditionally independent of w^*.

Theorem 4.4. *Suppose $d \ge 24$. Furthermore, suppose $n \ge \frac{(d-1)^2}{64M^2r^2}$ and $n \ge \frac{(d-1)^2}{64}$. For any data collection strategy for n data points and estimator $\hat{f}$ depending on those data points (and conditionally independent of the true label distribution), there exists a norm-M w^* such that,*

$$\mathbb{E}[err(\hat{f})] - err^* \ge \frac{1}{250000} err^* \frac{d}{n} \tag{235}$$

Proof. Since ψ is bounded by $1/4$ globally, from Lemma B.12:

$$I(Y_i; \pi^* | X_i) \leq M^2 \epsilon^2 r^2 \exp\left(2\sqrt{d-1}\epsilon M r\right) \tag{236}$$

From the data-processing inequality and adaptive mutual information tensorization,

$$\begin{aligned} I(\hat{f}, \pi^*) &\leq I(\{(X_i, Y_i)\}_{i=1}^n; \pi^*) && (237)\\ &\leq \sum_{i=1}^n I(Y_i; \pi^* | X_i) && (238)\\ &\leq n I_x(Y; w^*) && (239)\\ &\leq n M^2 \epsilon^2 r^2 \exp\left(2\sqrt{d-1}\epsilon M r\right) && (240) \end{aligned}$$

Let $\epsilon = \frac{1}{16}\frac{\sqrt{d-1}}{Mr\sqrt{n}}$. By assumption, n is large enough so $4(d-1)\epsilon^2 \leq 1$ and $2Mr\sqrt{d-1}\epsilon \leq 1$.

Then,

$$\begin{aligned} I(\hat{f}, \pi^*) &\leq n M^2 \frac{1}{256}\frac{d-1}{M^2 r^2 n} r^2 \exp\left(2\sqrt{d-1}\epsilon M r\right) && (241)\\ &\leq \frac{d-1}{64}\frac{e}{4} && (242)\\ &\leq \frac{d-1}{64} && (243) \end{aligned}$$

Thus, the conditions of Lemma B.8 are satisfied, so,

$$\begin{aligned} \mathbb{E}[\text{err}(\hat{f})] - \text{err}^* &\geq \frac{1}{808} M r \sqrt{d-1} \frac{1}{256}\frac{d-1}{M^2 r^2 n} && (244)\\ &= \frac{1}{206848}\frac{\sqrt{d-1}}{Mr}\frac{d-1}{n} && (245) \end{aligned}$$

Finally, using Lemma 4.2,

$$\begin{aligned} \mathbb{E}[\text{err}(\hat{f})] - \text{err}^* &\geq \frac{1}{206848}\frac{7}{8}\text{err}^*\frac{d-1}{d}\frac{d}{n} && (246)\\ &\geq \frac{1}{250000}\text{err}^*\frac{d}{n} && (247) \end{aligned}$$

Finally, since the expectation includes the draw of w^*, the worst-case w^* has at least as much excess error.

□

D. Random sampling lower bound

The setup here is the same as for the adaptive lower bound. We show that with constant probability, the randomly sampled points have low information and then apply the key lemma.

Recall $\psi(u) = \sigma(u)\sigma(-u)$.

Define $f(x) = \max_{w \in C(\mathcal{W})} \psi(w \cdot x)$.

Lemma D.1. *If $2Mr\sqrt{d-1}\epsilon \leq 1$, for X drawn uniformly at random from a radius r sphere,*

$$\frac{1}{2}err^* \leq \mathbb{E}[f(X)] \leq (2e)err^* \tag{248}$$

Proof. For the lower bound,

$$\begin{aligned}
\mathbb{E}[f(X)] &= \mathbb{E}\left[\max_{w \in C(\mathcal{W})} \psi(w \cdot X)\right] && (249)\\
&\geq \max_{w \in C(\mathcal{W})} \mathbb{E}[\psi(w \cdot X)] && (250)\\
&= \mathbb{E}[\psi(MX_1)] && (251)\\
&\geq \frac{1}{2}\mathbb{E}[\sigma(-M|X_1|)] && (252)\\
&= \frac{1}{2}\text{err}^* && (253)
\end{aligned}$$

For the upper bound,

$$\begin{aligned}
f(x) &\leq \max_{w \in C(\mathcal{W})} \psi(w \cdot x) && (254)\\
&\leq \max_{w \in C(\mathcal{W})} \exp(-|w \cdot x|) && (255)\\
&= \exp(-|Me_1 \cdot x|) \max_{w \in C(\mathcal{W})} \exp(|Me_1 \cdot x| - |w \cdot x|) && (256)\\
&\leq 2\sigma(-M|x_1|) \max_{w \in C(\mathcal{W})} \exp(|Me_1 \cdot x| - |w \cdot x|) && (257)\\
& && (258)
\end{aligned}$$

Next, note that if $w \in C(\mathcal{W})$ (and thus $\|w\| \leq M$),

$$\begin{aligned}
|Me_1 \cdot x| - |w \cdot x| &\leq |(Me_1 - w) \cdot x| && (259)\\
&\leq \|Me_1 - w\|\|x\| && (260)\\
&\leq Mr\angle(e_1, w) && (261)\\
&\leq 2Mr\sqrt{d-1}\epsilon && (262)\\
&\leq 1 && (263)
\end{aligned}$$

The second to last line follow from Lemma B.5 and the last line follows by assumption.

Thus,

$$\begin{aligned}
f(x) &\leq 2\sigma(-M|x_1|)\exp(1) && (264)\\
\mathbb{E}[f(X)] &\leq (2e)\mathbb{E}[\sigma(-M|X_1|)] && (265)\\
&\leq (2e)\text{err}^* && (266)
\end{aligned}$$

□

Lemma D.2. *If $2Mr\sqrt{d-1}\varepsilon \leq 1$ and $n\mathit{err}^* \geq 6\ln(2)$, then, with probability at least $\frac{1}{2}$ over a random sample of $\{X_i\}_{i=1}^n$,*

$$\sum_{i=1}^{n} I(Y_i;\pi^*|X_i=x_i) \leq 44\mathit{err}^* n M^2\epsilon^2 r^2 \exp(2\sqrt{d-1}\epsilon Mr) \tag{267}$$

$$\tag{268}$$

Proof. From Lemma B.12,

$$I(Y_i;\pi^*|X_i=x) \leq 4\left[\max_{w\in C(\mathcal{W})}\psi(w\cdot x)\right] M^2\epsilon^2 r^2 \exp\left(2\sqrt{d-1}\epsilon Mr\right) \tag{269}$$

So, for randomly sampled $\{x_i\}_{i=1}^n$,

$$\sum_{i=1}^{n} I(Y_i;\pi^*|X_i=x_i) \leq \left[\sum_{i=1}^{n}\max_{w\in C(\mathcal{W})}\psi(w\cdot x_i)\right]\cdot 4M^2\epsilon^2 r^2\exp(2\sqrt{d-1}\epsilon Mr) \tag{270}$$

Note that $\psi(u)\in[0,1]$ so $f(X)\in[0,1]$ and we can apply a Chernoff bound (over the random sample) and use D.1

$$\Pr\left(\sum_{i=1}^{n} f(X_i) \geq 2\mathbb{E}\left[\sum_{i=1}^{n} f(X_i)\right]\right) \leq \exp\left(-\frac{1}{3}\mathbb{E}\left[\sum_{i=1}^{n} f(X_i)\right]\right) \tag{271}$$

$$\Pr\left(\sum_{i=1}^{n} f(X_i) \geq (4e)\mathrm{err}^* n\right) \leq \exp\left(-\frac{1}{3}\frac{1}{2}\mathrm{err}^* n\right) \tag{272}$$

Then, if $n\mathrm{err}^* \geq 6\ln(2)$,

$$\Pr\left(\sum_{i=1}^{n} f(X_i) \geq (4e)\mathrm{err}^* n\right) \leq \frac{1}{2} \tag{273}$$

$$\Pr\left(\sum_{i=1}^{n} f(X_i) \leq (4e)\mathrm{err}^* n\right) \geq \frac{1}{2} \tag{274}$$

Thus, with probability at least $\frac{1}{2}$ over the random draw of the inputs,

$$\sum_{i=1}^{n} I(Y_i;\pi^*|X_i=x_i) \leq (4e)\mathrm{err}^* n\cdot 4M^2\epsilon^2 r^2\exp(2\sqrt{d-1}\epsilon Mr) \tag{275}$$

$$\tag{276}$$

□

Theorem 4.5. *Suppose $d\geq 24$. Furthermore, suppose $n\geq\frac{(d-1)^2}{64M^2r^2\cdot 44\mathit{err}^*}$, $n\geq\frac{(d-1)^2}{64\cdot 44\mathit{err}^*}$, and $n\geq\frac{6\ln(2)}{\mathit{err}^*}$. For any estimator $\hat{f}$ computed from n random samples (and conditionally independent of the true label distribution given the data), there exists a norm-M w^* such that:*

$$\mathbb{E}[\mathit{err}(\hat{f})] - \mathit{err}^* \geq \frac{1}{22000000}\frac{d}{n} \tag{277}$$

Proof. By Lemma D.2, with probability at least $1/2$ for a random draw of $\{x_i\}_{i=1}^n$,

$$\sum_{i=1}^{n} I(Y_i; \pi^* | X_i = x_i) \leq 44\text{err}^* n M^2 \epsilon^2 r^2 \exp(2\sqrt{d-1}\epsilon M r) \tag{278}$$

$$\tag{279}$$

From the data-processing inequality and mutual information tensorization,

$$\begin{aligned} I(\hat{f}, \pi^* | \{X_i = x_i\}_{i=1}^n) &\leq I(\{(X_i, Y_i)\}_{i=1}^n; \pi^* | \{X_i = x_i\}_{i=1}^n) && (280)\\ &\leq \sum_{i=1}^{n} I(Y_i; \pi^* | X_i = x_i) && (281)\\ &\leq 44\text{err}^* n M^2 \epsilon^2 r^2 \exp\left(2\sqrt{d-1}\epsilon M r\right) && (282) \end{aligned}$$

Let $\epsilon = \frac{1}{\sqrt{44\text{err}^*}} \frac{1}{16} \frac{\sqrt{d-1}}{Mr\sqrt{n}}$. Then, by the assumption on n, $n\text{err}^* \geq 6\ln(2)$, $2Mr\sqrt{d-1}\varepsilon \leq 1$, and $4(d-1)\varepsilon^2 \leq 1$

Then

$$\begin{aligned} I(\hat{f}, \pi^* | \{X_i = x_i\}_{i=1}^n) &\leq 44\text{err}^* n M^2 \frac{1}{256} \frac{d-1}{M^2 r^2 n} \frac{1}{44\text{err}^*} r^2 \exp\left(2\sqrt{d-1}\epsilon M r\right) && (283)\\ &\leq \frac{d-1}{64} \frac{e}{4} && (284)\\ &\leq \frac{d-1}{64} && (285) \end{aligned}$$

Thus, the conditions of Lemma B.8 are satisfied:

$$\begin{aligned} \mathbb{E}[\text{err}(\hat{f})] - \text{err}^* &\geq \frac{1}{808} Mr\sqrt{d-1} \frac{1}{256} \frac{d-1}{M^2 r^2 n} \frac{1}{44\text{err}^*} && (286)\\ &= \frac{1}{9101312\text{err}^*} \frac{\sqrt{d-1}}{Mr} \frac{d-1}{n} && (287) \end{aligned}$$

Finally, using Lemma 4.2,

$$\begin{aligned} \mathbb{E}[\text{err}(\hat{f})] - \text{err}^* &\geq \frac{1}{9101312\text{err}^*} \frac{7}{8} \text{err}^* \frac{d-1}{d} \frac{d}{n} && (288)\\ &\geq \frac{1}{11000000} \frac{d}{n} && (289) \end{aligned}$$

Thus, with probability $1/2$ over randomly drawn x, the expected excess loss is lower bounded. Since the excess loss is non-negative, the expectation (including the randomization over x) is lower bounded as,

$$\mathbb{E}[\text{err}(\hat{f})] - \text{err}^* \geq \frac{1}{22000000} \frac{d}{n} \tag{290}$$

Finally, since the expectation includes the draw of w^*, the worst-case w^* has at least as much excess error.

□

E. Random sampling upper bounds

The proof idea in this section is inspired by the proof of Theorem 5.1 in Frostig et al. (2015).

Let $\hat{L}_n(w)$ be the empirical logistic loss on n data points at parameter value w. Likewise, define $L(w)$ as the population logistic loss (with respect to a uniform distribution over a radius r sphere) at parameter value w.

For this section, define $Q = \nabla^2 L(w^*)$ (a form of the Fisher information (Lehmann & Casella, 2006)) and $\psi(u) = \sigma(u)\sigma(-u)$. Q will be featured prominently in this analysis, so first we find and bound it.

E.1. Calculation and bounds on Q

Without loss of generality, let $w^* = Me_1$. Let ξ be the first component of d-dimensional vector drawn uniformly from a sphere centered at the origin of radius Mr.

Lemma E.1. *Q is a diagonal matrix with*

$$Q_{1,1} = \frac{1}{M^2}\mathbb{E}[\psi(\xi)\xi^2] \tag{291}$$

and for any $i > 1$,

$$Q_{i,i} = \frac{1}{M^2}\left(\frac{M^2r^2}{d-1}\mathbb{E}[\psi(\xi)] - \frac{1}{d-1}\mathbb{E}[\psi(\xi)\xi^2]\right) \tag{292}$$

Proof. Note that

$$Q = \mathbb{E}[\psi(w^* \cdot x)xx^T] \tag{293}$$
$$Q_{i,j} = \mathbb{E}[\psi(Mx_1)x_ix_j] \tag{294}$$
$$\tag{295}$$

For the sphere, because of symmetry about the origin, note that $\mathbb{E}[x_j|x_i] = 0$ for $i \neq j$ and any value of x_i. Therefore, by the law of total expectation, $Q_{i,j} = 0$ for $i \neq j$ and thus Q is diagonal.

$$Q_{1,1} = \mathbb{E}[\psi(Mx_1)x_1^2] \tag{296}$$
$$= \frac{1}{M^2}\mathbb{E}[\psi(Mx_1)(Mx_1)^2] \tag{297}$$
$$= \frac{1}{M^2}\mathbb{E}[\psi(\xi)\xi^2] \tag{298}$$

Additionally, for $i > 1$

$$Q_{i,i} = \mathbb{E}[\psi(Mx_1)x_i^2] \tag{299}$$
$$= \mathbb{E}[\psi(Mx_1)\mathbb{E}[x_i^2|x_1]] \tag{300}$$

Note that after conditioning on x_1, the vector $x_{2:}$ is drawn uniformly from a $(d-1)$-dimensional sphere centered at the origin of radius $r^2 - x_1^2$. Therefore, conditioning on a value of x_1,

$$\sum_{i>1} x_i^2 = r^2 - x_1^2 \tag{301}$$

$$\sum_{i>1} \mathbb{E}[x_i^2|x_1] = r^2 - x_1^2 \tag{302}$$

$$(d-1)\mathbb{E}[x_i^2|x_1] = r^2 - x_1^2 \tag{303}$$

Therefore,

$$Q_{i,i} = \mathbb{E}\left(\psi(Mx_1)\frac{r^2 - x_1^2}{d-1}\right) \tag{304}$$

$$= \frac{r^2}{d-1}\mathbb{E}[\psi(Mx_1)] - \frac{1}{d-1}\mathbb{E}[\psi(Mx_1)x_1^2] \tag{305}$$

$$= \frac{1}{M^2}\left(\frac{M^2r^2}{d-1}\mathbb{E}[\psi(\xi)] - \frac{1}{d-1}\mathbb{E}[\psi(\xi)\xi^2]\right) \tag{306}$$

□

This lemma motivates the definition of $Q_1 = Q_{1,1}$ and $Q_2 = Q_{i,i}$ for $i > 1$.

E.1.1. BOUNDS ON EXPECTATIONS

Lemma E.2. *If* $d \geq 5$ *and* $\frac{\sqrt{d-1}}{Mr} \leq \frac{1}{6}$,

$$\mathbb{E}[\psi(\xi)\xi^2] \geq \frac{1}{32}\frac{\sqrt{d-1}}{Mr} \tag{307}$$

Proof. By assumption, $\frac{(d-3)6^2}{2M^2r^2} \leq \frac{1}{2}$. Then, by Lemma A.2, for $\alpha \leq 6$,

$$\frac{1}{5}\sqrt{d-1}\frac{\alpha}{Mr} \leq \Pr(|\xi| \leq \alpha) \leq \frac{8}{7}\sqrt{d-1}\frac{\alpha}{Mr} \tag{308}$$

Note that for any ξ,

$$\psi(\xi)\xi^2 \geq \frac{1}{20}\mathbf{1}[1/2 \leq |\xi| \leq 6] \tag{309}$$

So,

$$\mathbb{E}[\psi(\xi)\xi^2] \geq \frac{1}{20}\mathbb{E}[\mathbf{1}[1/2 \leq |\xi| \leq 6]] \tag{310}$$

$$= \frac{1}{20}\left(\Pr(|\xi| \leq 6) - \Pr(|\xi| \leq 1/2)\right) \tag{311}$$

$$\geq \frac{1}{20}\left(\frac{1}{5}\sqrt{d-1}\frac{6}{Mr} - \frac{8}{7}\sqrt{d-1}\frac{1/2}{Mr}\right) \tag{312}$$

$$= \frac{1}{20}\left(\frac{1}{5}6 - \frac{8}{7}\frac{1}{2}\right)\frac{\sqrt{d-1}}{Mr} \tag{313}$$

$$\geq \frac{1}{32}\frac{\sqrt{d-1}}{Mr} \tag{314}$$

□

Lemma E.3. *If $d \geq 4$,*

$$\mathbb{E}[\psi(\xi)\xi^2] \leq 12\frac{\sqrt{d-1}}{Mr} \tag{315}$$

Proof. Let us examine the following ratio:

$$\frac{\psi(\xi)\xi^2}{\sigma(-|\xi|/2)} = \frac{(1+\exp(|\xi|/2))|\xi|^2}{(1+\exp(|\xi|))(1+\exp(-|\xi|)} \tag{316}$$

$$\leq \frac{2\exp(|\xi|/2)}{\exp(|\xi|)}|\xi|^2 \tag{317}$$

$$= 2\exp(-|\xi|/2 + 2\ln(|\xi|)) \tag{318}$$

The expression $-|\xi|/2 + 2\ln(|\xi|)$ is maximized at $|\xi| = 4$, so

$$2\exp(-|\xi|/2 + 2\ln(\xi)) \leq \frac{32}{e^2} \leq 5 \tag{319}$$

Therefore, for all ξ,

$$\psi(\xi)\xi^2 \leq 5\sigma(-|\xi|/2) \tag{320}$$

Since ξ and Mx_1 have the same distribution:

$$\mathbb{E}[\psi(\xi)\xi^2] \leq 5\mathbb{E}[\sigma(-|\xi|/2)] \tag{321}$$

$$= 5\mathbb{E}[\sigma(-M|x_1|/2)] \tag{322}$$

$$= 5\text{err}^*(M/2) \tag{323}$$

where $\text{err}^*(M/2)$ is the error for parameters of norm $M/2$. From Lemma 4.2,

$$\text{err}^*(M/2) \leq \frac{8}{7}\frac{\sqrt{d-1}}{(M/2)r} \tag{324}$$

Putting these together, we get the result. □

Lemma E.4. *If $d \geq 5$ and $\frac{\sqrt{d-1}}{MR} \leq 1$,*

$$\frac{1}{60}\frac{\sqrt{d-1}}{Mr} \leq \mathbb{E}[\psi(\xi)] \leq \frac{8}{7}\frac{\sqrt{d-1}}{Mr} \tag{325}$$

Proof. Note that for any ξ,

$$\frac{1}{2}\sigma(-|\xi|) \leq \psi(\xi) \leq \sigma(-|\xi|) \tag{326}$$

Therefore, since ξ and Mx_1 have the same distribution:

$$\frac{1}{2}\mathbb{E}[\sigma(-M|x_1|)] \leq \mathbb{E}[\psi(\xi)] \leq \mathbb{E}[\sigma(-M|x_1|)] \tag{327}$$

Thus,

$$\frac{1}{2}\text{err}^* \leq \mathbb{E}[\psi(\xi)] \leq \text{err}^* \tag{328}$$

Using the assumption $\frac{\sqrt{d-1}}{MR} \leq 1$, Lemma 4.1, and Lemma 4.2.

$$\frac{1}{40}\frac{\sqrt{d-1}}{Mr} \leq \mathbb{E}[\psi(\xi)] \leq \frac{8}{7}\frac{\sqrt{d-1}}{Mr} \tag{329}$$

□

E.1.2. Bounds on Q

Lemma E.5. *If $d \geq 5$ and $\frac{\sqrt{d-1}}{Mr} \leq \frac{1}{12}$,*

$$Q_1 \geq \frac{1}{32}\frac{1}{M^2}\frac{\sqrt{d-1}}{Mr} \tag{330}$$

$$Q_2 \geq \frac{1}{240}\frac{1}{M^2}\frac{Mr}{\sqrt{d-1}} \tag{331}$$

$$\lambda_{min}(Q) \geq \frac{1}{32}\frac{1}{M^2}\frac{\sqrt{d-1}}{Mr} \tag{332}$$

Proof. From Lemma E.1 and Lemma E.2,

$$Q_1 \geq \frac{1}{32}\frac{1}{M^2}\frac{\sqrt{d-1}}{Mr} \tag{333}$$

From Lemma E.1, Lemma E.3, and Lemma E.4,

$$Q_2 \geq \frac{1}{M^2}\left(\frac{M^2r^2}{d-1}\frac{1}{40}\frac{\sqrt{d-1}}{Mr} - \frac{1}{d-1}12\frac{\sqrt{d-1}}{Mr}\right) \tag{334}$$

$$= \frac{1}{M^2}\frac{Mr}{\sqrt{d-1}}\left(\frac{1}{40} - \frac{12}{d-1}\frac{d-1}{M^2r^2}\right) \tag{335}$$

Using the assumptions on $\frac{\sqrt{d-1}}{Mr}$ and d

$$Q_2 \geq \frac{1}{M^2}\frac{Mr}{\sqrt{d-1}}\left(\frac{1}{40} - \frac{12}{4}\frac{1}{144}\right) \tag{336}$$

$$= \frac{1}{240}\frac{1}{M^2}\frac{Mr}{\sqrt{d-1}} \tag{337}$$

Finally,

$$\lambda_{\min}(Q) = \min(Q_1, Q_2) \tag{338}$$
$$\geq \frac{1}{32}\frac{1}{M^2}\frac{\sqrt{d-1}}{Mr} \tag{339}$$

Where the last line follows from $\sqrt{d-1}/Mr$ being small so the lower bound on Q_1 is lower than the lower bound on Q_2.

□

E.2. Geometric arguments

We define four regions around w^* with size defined by q: an ellipsoid, a ball, a cylinder, and a cone. Without loss of generality, assume $w^* = Me_1$. Let $w_{2:}$ denote the vector w without the first component.

$$R_{\text{Ellipsoid}} = \left\{ w : (w - w^*)^T Q (w - w^*) \leq q^2 \right\} \tag{340}$$
$$R_{\text{Ball}} = \left\{ w : \|w - w^*\| \leq \frac{q}{\sqrt{\lambda_{\min}(Q)}} \right\} \tag{341}$$
$$R_{\text{Cylinder}} = \left\{ w : \frac{M}{2} \leq w_1 \leq \frac{3M}{2}, \|w_{2:}\| \leq \frac{q}{\sqrt{Q_2}} \right\} \tag{342}$$
$$R_{\text{Cone}} = \left\{ w : \angle(w, w^*) \leq q\frac{2}{M\sqrt{Q_2}} \right\} \tag{343}$$

We show that $R_{\text{Ellipsoid}} \subset R_{\text{Ball}}$ and that, under some conditions, $R_{\text{Ellipsoid}} \subset R_{\text{Cylinder}} \subset R_{\text{Cone}}$.

Lemma E.6.

$$R_{\textit{Ellipsoid}} \subset R_{\textit{Ball}} \tag{344}$$

Proof. For any point $w \in R_{\text{Ellipsoid}}$,

$$(w - w^*)^T Q (w - w^*) \leq q^2 \tag{345}$$

and thus

$$\lambda_{\min}(Q)\|w - w^*\|^2 \leq q^2 \tag{346}$$
$$\|w - w^*\| \leq q\frac{1}{\sqrt{\lambda_{\min}(Q)}} \tag{347}$$

and thus $w \in R_{\text{Ball}}$. □

Lemma E.7. *If* $q \leq M\sqrt{Q_1}/2$,

$$R_{\textit{Ellipsoid}} \subset R_{\textit{Cylinder}} \tag{348}$$

Proof.

$$R_{\text{Ellipsoid}} = \{w : (w - w^*)^T Q (w - w^*) \leq q^2\} \tag{349}$$
$$= \{w : (w_1 - M)^2 Q_1 + \sum_{i>1} w_i^2 Q_i \leq q^2\} \tag{350}$$
$$= \{w : Q_1 (w_1 - M)^2 + Q_2 \|w_{2:}\|^2 \leq q^2\} \tag{351}$$

Since Q_1 and Q_2 are positive, any point within $R_{\text{Ellipsoid}}$ satisfies

$$Q_1 (w_1 - M)^2 \leq q^2 \tag{352}$$
$$Q_2 \|w_{2:}\|^2 \leq q^2 \tag{353}$$

Furthermore, if $q \leq M\sqrt{Q_1}/2$,

$$(w_1 - M)^2 \leq \left(\frac{M}{2}\right)^2 \tag{354}$$

so

$$\frac{M}{2} \leq w_1 \leq \frac{3M}{2} \tag{355}$$

□

Lemma E.8.

$$R_{\textit{Cylinder}} \subset R_{\textit{Cone}} \tag{356}$$

Proof. Suppose that $w \in R_{\text{Cylinder}}$.

Note that

$$\|w_{2:}\| = |\tan(\angle(w, e_1))| w_1 \tag{357}$$
$$\|w_{2:}\|^2 = \tan^2(\angle(w, w^*)) w_1^2 \tag{358}$$

using the properties of the definition of R_{Cylinder}, $\|w_{2:}\|^2 \leq q^2/Q_2$ and $w_1 \geq M/2$,

$$\frac{q^2}{Q_2} \geq \tan^2(\angle(w, w^*)) \frac{M^2}{4} \tag{359}$$

Noting that $\tan^2(u) \geq u^2$ for $u \leq \pi/2$ (also note $\angle(w, w^*) \leq \pi/2$ since $w_1 \geq M/2 > 0$),

$$\frac{q^2}{Q_2} \geq \angle(w, w^*)^2 \frac{M^2}{4} \tag{360}$$
$$\angle(w, w^*) \leq q \frac{2}{M\sqrt{Q_2}} \tag{361}$$

and thus $w \in R_{\text{Cone}}$. □

E.3. Hessian bounds

Lemma E.9. *If* $n \geq \ln(d/\delta)\frac{2r^2}{\lambda_{min}(Q)}$, *then with probability* $1-\delta$

$$2\nabla^2 \hat{L}_n(w^*) \succeq \nabla^2 L(w^*) \tag{362}$$

Proof. Note that

$$\nabla^2 \hat{L}_n(w^*) = \frac{1}{n}\sum_{i=1}^{n} \nabla^2 \ell(x_i, y_i, w^*) \tag{363}$$

$$n\nabla^2 \hat{L}_n(w^*) = \sum_{i=1}^{n} \nabla^2 \ell(x_i, y_i, w^*) \tag{364}$$

In the notation of Theorem 5.1.1 of Tropp (2015) (a matrix Chernoff bound),

Let Y be defined as,

$$Y = nQ^{-1/2}\nabla^2 \hat{L}_n(w^*)Q^{-1/2} \tag{365}$$

$$= \sum_{i=1}^{n} Q^{-1/2}\nabla^2 \ell(x_i, y_i, w^*)Q^{-1/2} \tag{366}$$

By convexity,

$$\lambda_{\min}\left(Q^{-1/2}\nabla^2 \ell(x_i, y_i, w^*)Q^{-1/2}\right) \geq 0 \tag{367}$$

Furthermore, since $\nabla^2 \ell(x_i, y_i, w^*) = \psi(w^* \cdot x_i)x_i x_i^T$ and since $\psi(u)$ is bounded by $1/4$,

$$\lambda_{\max}\left(Q^{-1/2}\nabla^2 \ell(x_i, y_i, w^*)Q^{-1/2}\right) \leq \frac{r^2}{4}\frac{1}{\lambda_{\min}(Q)} = L \tag{368}$$

Finally, $\mathbb{E}[Y] = nI_d$. Thus, with $\varepsilon = 1/2$ and Theorem 5.1.1 of Tropp (2015),

$$\Pr(\lambda_{\min}(Y) \leq n/2) \leq d\left[\frac{e^{-1/2}}{\sqrt{1/2}}\right]^{n/L} \tag{369}$$

$$\leq d\exp\left(-\frac{1}{8}\right)^{n/L} \tag{370}$$

$$\leq d\exp\left(-\frac{\lambda_{\min}(Q)n}{2r^2}\right) \tag{371}$$

So, if $n \geq \ln(d/\delta)\frac{2r^2}{\lambda_{\min}(Q)}$ with probability at least $1-\delta$,

$$\lambda_{\min}(Q^{-1/2}\nabla^2 \hat{L}_n(w^*)Q^{-1/2}) \geq \frac{1}{2} \tag{372}$$

$$Q^{-1/2}\nabla^2 \hat{L}_n(w^*)Q^{-1/2} \succeq \frac{1}{2}I_d \tag{373}$$

$$\nabla^2 \hat{L}_n(w^*) \succeq \frac{1}{2}Q \tag{374}$$

$$2\nabla^2 \hat{L}_n(w^*) \succeq \nabla^2 L(w^*) \tag{375}$$

□

Lemma E.10. *If* $q \leq \frac{10}{4r^3}\left(\lambda_{min}(Q)\right)^{3/2}$ *and* $2\nabla^2\hat{L}_n(w^*) \succeq \nabla^2 L(w^*)$*, then for* $w \in R_{Ball}$*,*

$$2\nabla^2\hat{L}_n(w) \succeq \nabla^2\hat{L}_n(w^*) \tag{376}$$

Proof. First, note that $|\psi'(u)| \leq 1/10$. Fix any $w \in R_{\text{Ball}}$.

We will now show $\|\nabla^2\hat{L}_n(w) - \nabla^2\hat{L}_n(w^*)\| \leq qr^3/10$.

$$\begin{aligned}
\|\nabla^2\hat{L}_n(w) - \nabla^2\hat{L}_n(w^*)\| &= \left\|\frac{1}{n}\sum_{i=1}^n \psi(w^*\cdot x_i)x_ix_i^T - \frac{1}{n}\sum_{i=1}^n \psi(w\cdot x_i)x_ix_i^T\right\| && (377)\\
&\leq \max_{\|x\|=r}\left\|[\psi(w^*\cdot x) - \psi(w\cdot x)]xx^T\right\| && (378)\\
&= \max_{\|x\|=r,\|v\|=1} v^T\left[[\psi(w^*\cdot x) - \psi(w\cdot x)]xx^T\right]v && (379)\\
&= \max_{\|x\|=r,\|v\|=1} \psi'(\widetilde{w}\cdot x)x^T(w^*-w)(v\cdot x)^2 && (380)\\
&\leq \frac{r^3}{10}\|w-w^*\| && (381)
\end{aligned}$$

where $\widetilde{w}$ in the second-to-last line is some point between w^* and w.

So, for $w \in R_{\text{Ball}}$, $\|\nabla^2\hat{L}_n(w) - \nabla^2\hat{L}_n(w^*)\| \leq \frac{r^3}{10}q\frac{1}{\sqrt{\lambda_{\min}(Q)}} \leq \frac{1}{4}\lambda_{\min}(Q)$.

Therefore,

$$\begin{aligned}
\nabla^2\hat{L}_n(w) - \nabla^2\hat{L}_n(w^*) &\succeq -\frac{1}{4}Q && (382)\\
2\nabla^2\hat{L}_n(w) - 2\nabla^2\hat{L}_n(w^*) &\succeq -\frac{1}{2}Q && (383)\\
2\nabla^2\hat{L}_n(w) - \nabla^2\hat{L}_n(w^*) &\succeq \nabla^2\hat{L}_n(w^*) - \frac{1}{2}Q && (384)\\
&\succeq \frac{1}{2}\left(2\nabla^2\hat{L}_n(w^*) - Q\right) && (385)\\
&\succeq 0 && (386)
\end{aligned}$$

□

E.4. High probability bound on gradient of empirical loss

Define $Z = Q^{-1/2}\nabla\hat{L}_n(w^*)$.

Lemma E.11. *If* $n \geq \frac{4}{9}\frac{r^2}{\lambda_{min}(Q)}\ln(2d/\delta)$*, then with probability* $1-\delta$*,*

$$\|Z\|^2 \leq \frac{4d\ln(2d/\delta)}{n} \tag{387}$$

Proof.

$$\begin{aligned}
Z &= Q^{-1/2}\nabla\hat{L}_n(w^*) && (388)\\
&= \frac{1}{n}\sum_{i=1}^n Q^{-1/2}\nabla\ell(x_i,y_i,w^*) && (389)
\end{aligned}$$

Define $Z^{(i)} = Q^{-1/2}\nabla\ell(x_i, y_i, w^*)$

Note that $\|Z^{(i)}\| \le \|Q^{-1/2}\|\|\nabla\ell(x_i, y_i, w^*)\| \le \frac{r}{\sqrt{\lambda_{\min}(Q)}}$ and thus $Z_j^{(i)} \le \frac{r}{\sqrt{\lambda_{\min}(Q)}}$.

Further note that

$$\mathbb{E}[Z^{(i)}] = 0 \tag{390}$$
$$\mathrm{Cov}(Z^{(i)}) = Q^{-1/2}\mathbb{E}[\nabla\ell(x, y, w^*)\nabla\ell(x, y, w^*)^T]Q^{-1/2} \tag{391}$$
$$= I_d \tag{392}$$

The above follows from noting that $\mathbb{E}[\sigma(-yx \cdot w^*)^2 xx^T] = \mathbb{E}[\psi(x \cdot w^*)xx^T] = Q$, a standard identity for the Fisher information.

By Bernstein's inequality, for any dimension j,

$$\Pr\left(|nZ_j| \ge \sqrt{4n\ln(2d/\delta)}\right) \le 2\exp\left(-\frac{\frac{1}{2}4n\ln(2d/\delta)}{n + \frac{1}{3}\frac{r}{\sqrt{\lambda_{\min}(Q)}}\sqrt{4n\ln(2d/\delta)}}\right) \tag{393}$$

$$\Pr\left(Z_j^2 \ge \frac{4\ln(2d/\delta)}{n}\right) \le 2\exp\left(-\frac{2\ln(2d/\delta)}{1 + \frac{2}{3}\frac{r}{\sqrt{\lambda_{\min}(Q)}}\sqrt{\frac{\ln(2d/\delta)}{n}}}\right) \tag{394}$$

If $n \ge \frac{4}{9}\frac{r^2}{\lambda_{\min}(Q)}\ln(2d/\delta)$, then

$$\Pr\left(Z_j^2 \ge \frac{4\ln(2d/\delta)}{n}\right) \le 2\exp\left(-\frac{2\ln(2d/\delta)}{2}\right) \tag{395}$$
$$= \delta/d \tag{396}$$

Then, by a union bound over all dimensions, with probability $1 - \delta$,

$$\forall j : Z_j^2 \le \frac{4\ln(2d/\delta)}{n} \tag{397}$$
$$\|Z\|^2 \le \frac{4d\ln(2d/\delta)}{n} \tag{398}$$

□

E.5. Main argument

Theorem E.12. *Suppose $d \ge 5$ and $\frac{\sqrt{d-1}}{Mr} \le 1/12$. Let $\hat{w}$ be the logistic MLE estimator from n randomly sampled points on a radius r sphere. For any $\delta > 0$, if $n \ge 64r^3M^3\ln(4d/\delta)/\sqrt{d-1}$, $n \ge 1600000M^9r^9d\ln(4d/\delta)/(d-1)^{3/2}$, and $n \ge 18500d\ln(4d/\delta)Mr/\sqrt{d-1}$, then with probability $1 - \delta$,*

$$\angle(\hat{w}, w^*)^2 \le 300000\frac{\sqrt{d-1}}{Mr}\frac{d}{n}\ln(4d/\delta) \tag{399}$$

and

$$\frac{M}{2} \leq \|\hat{w}\| \leq \frac{3M}{2} \tag{400}$$

Proof. Set $q = 17\sqrt{\frac{d}{n}\ln(4d/\delta)}$ so that $q^2 > 256\frac{d}{n}\ln(4d/\delta)$.

By a Taylor expansion,

$$\hat{L}_n(w) = \hat{L}_n(w^*) + \left[\nabla \hat{L}_n(w^*)\right]^T (w - w^*) + \frac{1}{2}(w - w^*)^T \left[\nabla^2 \hat{L}_n(\widetilde{w})\right](w - w^*) \tag{401}$$

for some $\widetilde{w}$ between w and w^*.

By assumption and by Lemma E.5, $n \geq 64r^3M^3\ln(4d/\delta)/\sqrt{d-1} \geq \ln(4d/\delta)\frac{2r^2}{\lambda_{\min}(Q)}$.

Thus, by Lemma E.9, with probability at least $1 - \delta/2$, $\nabla^2 \hat{L}_n(w^*) \succeq \frac{1}{2}\nabla^2 L(w^*)$.

By assumption and Lemma E.5, $n \geq 1600000M^9r^9d\ln(4d/\delta)/(d-1)^{3/2}$, so $q \leq \frac{10}{4r^3}\left(\lambda_{\min}(Q)\right)^{3/2}$

By Lemma E.10, if $\nabla^2 \hat{L}_n(w) \succeq \frac{1}{2}\nabla^2 \hat{L}_n(w^*)$ (which occurs with probability at least $1 - \delta/2$), for any $w \in R_{\text{Ball}}$ (and also for $w \in R_{\text{Ellipsoid}}$ by Lemma E.6),

$$\hat{L}_n(w) \geq \hat{L}_n(w^*) + \left[\nabla \hat{L}_n(w^*)\right]^T (w - w^*) + \frac{1}{8}(w - w^*)^T \left[\nabla^2 L(w^*)\right](w - w^*) \tag{402}$$

$$\tag{403}$$

Then, for w on the boundary of $R_{\text{Ellipsoid}}$, the following is true,

$$\hat{L}_n(w) \geq \hat{L}_n(w^*) + \left(Q^{-1/2}\nabla \hat{L}_n(w^*)\right)^T Q^{1/2}(w - w^*) + \frac{1}{8}q^2 \tag{404}$$

$$\hat{L}_n(w) \geq \hat{L}_n(w^*) - \|Z\|q + \frac{1}{8}q^2 \tag{405}$$

By assumption and Lemma E.5, $n \geq 64r^3M^3\ln(4d/\delta)/\sqrt{d-1} \geq \frac{4}{9}r^2\ln(4d/\delta)/\lambda_{\min}(Q)$.

Thus, by Lemma E.11, with probability $1 - \delta/2$, $\|Z\| \leq \sqrt{\frac{4d\ln(4d/\delta)}{n}} < \frac{1}{8}q$, by the definition of q.

Then, with probability $1 - \delta$ (union bound over the two $\delta/2$ events), for all w on the boundary of $R_{\text{Ellipsoid}}$,

$$\hat{L}_n(w) > \hat{L}_n(w^*) \tag{406}$$

Then, by convexity of $\hat{L}_n$, $R_{\text{Ellipsoid}}$ must contain the minimizer of $\hat{L}_n$, which we refer to as $\hat{w}$.

By assumption and Lemma E.5, $n \geq 18500d\ln(4d/\delta)Mr/\sqrt{d-1}$ so $q \leq M\sqrt{Q_1}/2$.

Therefore, by Lemmas E.7 and E.8, $R_{\text{Ellipsoid}} \subset R_{\text{Cylinder}} \subset R_{\text{Cone}}$. And thus, $\hat{w} \in R_{\text{Cone}}$. (Also note that since $\hat{w} \in R_{\text{Cylinder}}$, $M/2 \leq \|\hat{w}\| \leq 3M/2$.)

By the definition of R_{Cone},

$$\angle(\hat{w}, w^*) \leq q\frac{2}{M\sqrt{Q_2}} \tag{407}$$

$$\angle(\hat{w}, w^*)^2 \leq q^2\frac{4}{M^2Q_2} \tag{408}$$

$$\leq 289\frac{d}{n}\ln(4d/\delta)\frac{4}{M^2Q_2} \tag{409}$$

$$\leq 289\frac{d}{n}\ln(4d/\delta)\frac{4\cdot 240\sqrt{d-1}}{Mr} \tag{410}$$

$$\leq 300000\frac{\sqrt{d-1}}{Mr}\frac{d}{n}\ln(4d/\delta) \tag{411}$$

□

Lemma E.13. *Let a and b be positive real numbers, and $\delta_{min} \in [0,1]$. If $G \in [0,1]$ is a random variable and for $\delta \in [\delta_{min}, 1]$*

$$\Pr(G \leq a + b\ln(1/\delta)) \geq 1 - \delta \tag{412}$$

then

$$\mathbb{E}[G] \leq a + b + \delta_{min} \tag{413}$$

Proof. Note that by assumption, for $\delta \in [\delta_{\min}, 1]$,

$$\Pr(G \geq a + b\ln(1/\delta)) \leq \delta \tag{414}$$

Rearranging,

$$\Pr(G \geq g) \leq \exp\left(-\frac{g-a}{b}\right) \tag{415}$$

We can use this bound up until $g = a + b\ln(1/\delta_{\min})$. Call this upper limit $U = a + b\ln(1/\delta_{\min})$.

Since G is a non-negative random variable,

$$\mathbb{E}[G] = \int_0^\infty \Pr(G \geq g)dg \tag{416}$$

We now look at two cases. Suppose $U \leq 1$,

$$\mathbb{E}[G] = \int_0^a \Pr(G \geq g)dg + \int_a^U \Pr(G \geq g)dg + \int_U^1 \Pr(G \geq g)dg + \int_1^\infty \Pr(G \geq g)dg \tag{417}$$

For the first integral, note that probabilities of events are upper bounded by 1.

For the second integral, we use the assumed bound.

For the third integral, we note that the inverse cdf of a random variable is decreasing.

For the fourth integral, we note that G is upper bounded by 1.

$$\mathbb{E}[G] \leq \int_0^a 1dg + \int_a^U \exp\left(-\frac{g-a}{b}\right) dg + (1-U)\Pr(G \geq U) + 0 \tag{418}$$

$$\leq a + \int_a^\infty \exp\left(-\frac{g-a}{b}\right) dg + \Pr(G \geq U) \tag{419}$$

For the second case, suppose $U \geq 1$ and use a similar bounding strategy:

$$\mathbb{E}[G] = \int_0^a \Pr(G \geq g)dg + \int_a^1 \Pr(G \geq g)dg + \int_1^\infty \Pr(G \geq g)dg \tag{420}$$

$$\leq \int_0^a 1dg + \int_a^1 \exp\left(-\frac{g-a}{b}\right) dg + 0 \tag{421}$$

$$\leq a + \int_a^\infty \exp\left(-\frac{g-a}{b}\right) dg \tag{422}$$

So in either case,

$$\mathbb{E}[G] \leq a + \int_a^\infty \exp\left(-\frac{g-a}{b}\right) dg + \Pr(G \geq U) \tag{423}$$

$$\leq a + b + \exp\left(-\frac{U-a}{b}\right) \tag{424}$$

$$= a + b + \delta_{\text{min}} \tag{425}$$

□

Theorem 4.6. *Suppose $d \geq 5$ and $\frac{\sqrt{d-1}}{Mr} \leq 1/12$. Let $\hat{w}$ be the logistic MLE estimator from n randomly sampled points on a radius r sphere. If $n \geq \left(64r^3M^3/\sqrt{d-1}\right)^2$, $n \geq \left(1600000M^9r^9d/(d-1)^{3/2}\right)^2$, and $n \geq \left(18500dMr/\sqrt{d-1}\right)^2$, then*

$$\mathbb{E}[err(\hat{w})] - err(w^*) \leq 240000\frac{d\ln(d)}{n} \tag{426}$$

Proof. By Theorem E.12, for $\delta > 0$ and n sufficiently large (in terms of δ), the following holds with $1-\delta$ probability:

$$\angle(\hat{w}, w^*)^2 \leq 300000\frac{\sqrt{d-1}}{Mr}\frac{d}{n}\ln(4d/\delta) \tag{427}$$

which implies, via Lemma 4.3, that

$$\text{err}(\hat{w}) - \text{err}^* \leq 60000\frac{d}{n}\ln(4d/\delta) \tag{428}$$

We now apply Lemma E.13. Let $G = \text{err}(\hat{w}) - \text{err}^*$. Note that $G \geq 0$ by optimality of err^* and $G \leq 1$ since $\text{err}(w) \leq 1$ for all w.

Define $a = 60000\frac{d}{n}\ln(4d)$ and $b = 60000\frac{d}{n}$. Let $\delta_{\min} = 4d\exp(-\sqrt{n})$, note for later that $\delta_{\min} \le b$ for $n \ge 1$.

Noting that $n \ge 64r^3M^3\sqrt{n}/\sqrt{d-1} = 64r^3M^3\ln(4d/\delta_{\min})/\sqrt{d-1}$,

$n \ge 1600000M^9r^9d\sqrt{n}/(d-1)^{3/2} = 1600000M^9r^9d\ln(4d/\delta_{\min})/(d-1)^{3/2}$,

and $n \ge 18500d\sqrt{n}Mr/\sqrt{d-1} = 18500d\ln(4d/\delta_{\min})Mr/\sqrt{d-1}$,

therefore, for $\delta \in [\delta_{\min}, 1]$,

$$\Pr(G \ge a + b\ln(1/\delta)) \le \delta \tag{429}$$

by Lemma E.13,

$$\begin{aligned}
\mathbb{E}[\text{err}(\hat{w})] - \text{err}^* &\le a + b + \delta_{\min} && (430)\\
&\le a + 2b && (431)\\
&= 60000\frac{d}{n}(\ln(4d) + 2) && (432)\\
&\le 240000\frac{d\ln(d)}{n} && (433)
\end{aligned}$$

the last line follows from noticing $\ln(4d) + 2 \le 4\ln(d)$ for $d \ge 5$. □

F. Adaptive upper bound

F.1. Algorithm

Recall that after randomly sampling with half the budget and using the logistic MLE to find w_{random} we set $w_1 = \frac{2}{3}w_{\text{random}}$. Define $\hat{M} = \|w_1\|$. Then, $\mathcal{W} = \left\{w : \|w\| \le \hat{M}, \angle(w, w_1) \le \frac{1}{2}\min\left(\frac{\pi}{2}, \frac{2}{\hat{M}r}\right)\right\}$. See Algorithm 1 for more details. Define $S(w) = \{x : \|x\| = r, x \cdot w = 0\}$ as the decision boundary for weights w. We have the following iterates:

$$\begin{aligned}
x_t &\sim U(\{x : \|x\| = r, x \cdot w_t = 0\}) = U(S(w_t)) && (434)\\
y_t &\sim 2\text{Bernoulli}(\sigma(w^* \cdot x_t)) - 1 && (435)\\
g_t &= \nabla_w \ell(x_t, y_t, w_t) && (436)\\
&= -\sigma(x_t \cdot w_t)y_t x_t && (437)\\
&= -\frac{1}{2}y_t x_t && (438)\\
w_{t+1} &= \Pi_{\mathcal{W}}(w_t - \eta_t g_t) && (439)
\end{aligned}$$

F.2. Strong convexity

Note that $\frac{\|w\|}{M}w^*$ is the re-scaled w^* to have the same norm as w.

Lemma F.1. *For all w where $\angle(w, w^*) \le \min\left(\frac{\pi}{2}, \frac{2}{Mr}\right)$. If x is sampled uniformly from $S(w)$ and y is the corresponding sampled label (according to w^*),*

$$\mathbb{E}[\nabla\ell(x, y, w)] \cdot \left(w - \frac{\|w\|}{M}w^*\right) \ge \frac{1}{6}\frac{M}{\|w\|}\frac{r^2}{d-1}\left\|w - \frac{\|w\|}{M}w^*\right\|^2 \tag{440}$$

Proof.

$$\mathbb{E}[\nabla \ell(x,y,w)] \cdot \left(w - \frac{\|w\|}{M} w^*\right) = \mathbb{E}\left[-\frac{1}{2} yx\right] \cdot \left(w - \frac{\|w\|}{M} w^*\right) \tag{441}$$

$$= \frac{1}{2}\mathbb{E}\left[-yx \cdot w + yx \cdot \frac{\|w\|}{M} w^*\right] \tag{442}$$

$$= \frac{\|w\|}{2M}\mathbb{E}[yx \cdot w^*] \tag{443}$$

$$= \frac{\|w\|}{2M}\mathbb{E}[(\sigma(x \cdot w^*) - \sigma(-x \cdot w^*))x \cdot w^*] \tag{444}$$

Define $f(u) = (\sigma(u) - \sigma(-u))u$

$$\mathbb{E}[\nabla \ell(x,y,w)] \cdot \left(w - \frac{\|w\|}{M} w^*\right) = \frac{\|w\|}{2M}\mathbb{E}_{x \sim S(w)}[f(x \cdot w^*)] \tag{445}$$

Note that since $\angle(x,w) = \pi/2$,

$$|w^* \cdot x| = \|x\|\|w^*\||\cos(\angle(x,w^*))| \tag{446}$$

$$= rM|\sin(\angle(x,w) - \angle(x,w^*))| \tag{447}$$

$$\leq rM|\angle(x,w) - \angle(x,w^*)| \tag{448}$$

$$\leq rM\angle(w,w^*) \tag{449}$$

where the last line follows from the reverse triangle inequality.

Further note that $|u| \leq 2 \implies f(u) \geq \frac{u^2}{3}$

Therefore, if $rM\angle(w,w^*) \leq 2$,

$$\mathbb{E}[\nabla \ell(x,y,w)] \cdot \left(w - \frac{\|w\|}{M} w^*\right) \geq \frac{\|w\|}{2M}\mathbb{E}_{x \sim S(w)}\left[\frac{(x \cdot w^*)^2}{3}\right] \tag{450}$$

$$= \frac{\|w\|}{6M}\mathbb{E}_{x \sim S(w)}[(x \cdot w^*)^2] \tag{451}$$

Without loss of generality, assume w is in the same direction as e_d, and $w^* = M\cos(\angle(w,w^*))e_d + M\sin(\angle(w,w^*))e_1$

$$\mathbb{E}_{x \sim S(w)}\left[(x \cdot w^*)^2\right] = \mathbb{E}_{x \sim U(rS^{d-2})}[M^2 \sin^2(\angle(w,w^*))x_1^2] \tag{452}$$

$$= r^2 M^2 \sin^2(\angle(w,w^*))\mathbb{E}_{x \sim U(S^{d-2})}[x_1^2] \tag{453}$$

By symmetry of the sphere $S^{d-2} \subset \mathbb{R}^{d-1}$, $\mathbb{E}_{x \sim U(S^{d-2})}[x_1^2] = \frac{1}{d-1}$ so

$$\mathbb{E}_{x \sim S(w)}\left[(x \cdot w^*)^2\right] = \frac{r^2 M^2}{d-1}\sin^2(\angle(w,w^*)) \tag{454}$$

Now, we need to connect $\sin(\angle(w,w^*))$ back to $\|w - w^*\|$.

By the equation for a chord on a circle (and noting $\|w\| = \left\|\frac{\|w\|}{M} w*\right\|$)

$$\left\| w - \frac{\|w\|}{M} w* \right\|^2 = \|w\|^2 (2 - 2\cos(\angle(w, w^*))) \tag{455}$$

In general, for acute angles $\theta \leq \frac{\pi}{2}$, $2 - 2\cos(\theta) \leq 2\sin^2(\theta)$. Therefore,

$$\left\| w - \frac{\|w\|}{M} w* \right\|^2 \leq 2\|w\|^2 \sin^2(\angle(w, w^*)) \tag{456}$$

or, rearranging,

$$\sin^2(\angle(w, w^*)) \geq \frac{1}{2\|w\|^2} \left\| w - \frac{\|w\|}{M} w* \right\|^2 \tag{457}$$

Putting it all together, we arrive at:

$$\mathbb{E}[\nabla \ell(x, y, w)] \cdot \left(w - \frac{\|w\|}{M} w^* \right) \geq \frac{1}{6} \frac{M}{\|w\|} \frac{r^2}{d-1} \left\| w - \frac{\|w\|}{M} w^* \right\|^2 \tag{458}$$

□

F.3. Optimization argument

This argument is inspired by the proof of Lemma 1 in Rakhlin et al. (2012), which is from Nemirovski et al. (2009).

Lemma F.2. *Suppose we have a convex set $\mathcal{W}$, a reachable set $\mathcal{R} \subset \mathcal{W}$, an initialization $w_1 \in \mathcal{R}$ and a random "stochastic gradient" $g(w)$ that is a function of w and has a bounded expectation: $\forall w \in \mathcal{R} : \mathbb{E}[\|g(w)\|^2] \leq G$.*

If there exists $\lambda > 0$ and $\overline{w} \in \mathcal{W}$ such that $\mathbb{E}[g(w)] \cdot (w - \overline{w}) \geq \lambda \|w - \overline{w}\|^2$ for all $w \in \mathcal{R}$, then, for a orthogonal projected stochastic gradient update rule $w_{t+1} = \Pi_{\mathcal{W}}\left(w_t - \frac{1}{\lambda t} g(w_t)\right)$, if the iterates w_t always stay in $\mathcal{R}$, then for any $t \geq 3$,

$$\mathbb{E}[\|w_t - \overline{w}\|^2] \leq \frac{G^2}{\lambda^2 t} \tag{459}$$

Proof. Define $\eta_t = \frac{1}{\lambda t}$ to be the step size.

Note that since $\mathcal{W}$ is convex, $\overline{w} \in \mathcal{W}$, and orthogonal projections onto convex sets contract distances,

$$\begin{aligned}
\mathbb{E}\left[\|w_{t+1} - \overline{w}\|^2\right] &= \mathbb{E}\left[\|\Pi_{\mathcal{W}}(w_t - \eta_t g(w_t)) - \overline{w}\|^2\right] && (460)\\
&\leq \mathbb{E}\left[\|w_t - \eta_t g(w_t) - \overline{w}\|^2\right] && (461)\\
&= \mathbb{E}[\|w_t - \overline{w}\|^2] - 2\eta_t \mathbb{E}[g(w_t) \cdot (w_t - \overline{w})] + \eta_t^2 \mathbb{E}[\|g(w_t)\|^2] && (462)\\
&= \mathbb{E}[\|w_t - \overline{w}\|^2] - 2\eta_t \mathbb{E}\left[\mathbb{E}[g(w_t)] \cdot (w_t - \overline{w})\right] + \eta_t^2 \mathbb{E}[\|g(w_t)\|^2] && (463)\\
&\leq \mathbb{E}[\|w_t - \overline{w}\|^2] - 2\eta_t \lambda \mathbb{E}[\|w_t - \overline{w}\|^2] + \eta_t^2 G^2 && (464)\\
&= (1 - 2\eta_t \lambda)\mathbb{E}[\|w_t - \overline{w}\|^2] + \eta_t^2 G^2 && (465)
\end{aligned}$$

Plugging in the step size $\eta_t = \frac{1}{\lambda t}$:

$$\mathbb{E}[\|w_{t+1} - \overline{w}\|^2] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}[\|w_t - \overline{w}\|^2] + \frac{G^2}{\lambda^2 t^2} \tag{466}$$

Then, from the above equation with $t = 2$,

$$\mathbb{E}[\|w_{t+1} - \overline{w}\|^2] \leq 0 + \frac{G^2}{4\lambda^2} \tag{467}$$

$$\leq \frac{G^2}{\lambda^2(t+1)} \tag{468}$$

we proceed by induction for $t \geq 3$,

$$\mathbb{E}[\|w_{t+1} - \overline{w}\|^2] \leq \left(1 - \frac{2}{t}\right) \mathbb{E}[\|w_t - \overline{w}\|^2] + \frac{G^2}{\lambda^2 t^2} \tag{469}$$

$$\leq \left(1 - \frac{2}{t}\right) \frac{G^2}{\lambda^2 t} + \frac{G^2}{\lambda^2 t^2} \tag{470}$$

$$= \frac{G^2}{\lambda^2} \left(\frac{1}{t} - \frac{2}{t^2} + \frac{1}{t^2}\right) \tag{471}$$

$$= \frac{G^2}{\lambda^2} \frac{t-1}{t^2} \tag{472}$$

$$\leq \frac{G^2}{\lambda^2(t+1)} \tag{473}$$

And thus, the following is proven,

$$\mathbb{E}[\|w_t - \overline{w}\|^2] \leq \frac{G^2}{\lambda^2 t} \tag{474}$$

□

F.4. Connection between distance and angle

Lemma F.3. *For any vectors $u, v \in \mathbb{R}^d - \{0\}$,*

$$\angle(u, v) \leq \frac{2\pi}{\|v\|} \|u - v\| \tag{475}$$

Proof. Without loss of generality, let $v = \|v\| e_1$ and let $u = a e_1 + b e_2$.

Next, we split into two cases,

Case 1: $\|u - v\| \leq \frac{\|v\|}{2}$

In this case, $a \geq \frac{1}{2}\|v\|$ and thus, $1 \leq 2\frac{a}{\|v\|}$.

Then,

$$\angle(u,v) = \left|\arctan\left(\frac{b}{a}\right)\right| \tag{476}$$
$$\leq \left|\frac{b}{a}\right| \tag{477}$$
$$= \sqrt{\left(\frac{b}{a}\right)^2} \tag{478}$$
$$\leq \sqrt{4\left(\frac{a}{\|v\|}\right)^2\left(\frac{b}{a}\right)^2} \tag{479}$$
$$= 2\sqrt{\left(\frac{b}{\|v\|}\right)^2} \tag{480}$$
$$\leq 2\pi\sqrt{\left(1-\frac{a}{\|v\|}\right)^2+\left(\frac{b}{\|v\|}\right)^2} \tag{481}$$
$$= \frac{2\pi}{\|v\|}\sqrt{(\|v\|-a)^2+b^2} \tag{482}$$
$$= \frac{2\pi}{\|v\|}\|v-u\| \tag{483}$$

Case 2: $\|u-v\| \geq \frac{\|v\|}{2}$

$$\angle(u,v) \leq \pi \tag{484}$$
$$= \frac{2\pi}{\|v\|}\frac{\|v\|}{2} \tag{485}$$
$$\leq \frac{2\pi}{\|v\|}\|u-v\| \tag{486}$$

□

Let $G^2 = \frac{r^2}{4}$ so that $\|g(w_t)\|^2 = \frac{1}{4}\|x_t\|^2 = G^2$

F.5. Putting it together

Theorem 4.7. *Suppose $d \geq 5$ and $\frac{\sqrt{d-1}}{Mr} \leq 1/12$. If $n \geq 4$ is large enough so that*

$$n \geq 64r^3M^3\ln(4n/\mathit{err}^*)/\sqrt{d-1} \tag{487}$$
$$n \geq 1600000M^9r^9d\ln(4n/\mathit{err}^*)/(d-1)^{3/2} \tag{488}$$
$$n \geq 18500d\ln(4n/\mathit{err}^*)Mr/\sqrt{d-1} \tag{489}$$

and

$$400\frac{\sqrt{d-1}}{Mr}\frac{d}{n}\ln(4n/\mathit{err}^*) \leq \left(\frac{1}{2}\min\left(\frac{\pi}{2},\frac{2}{3Mr}\right)\right)^2 \tag{490}$$

then, for the estimator $\hat{w}$ returned from Algorithm 1,

$$\mathbb{E}[\mathit{err}(\hat{w})] - \mathit{err}(w^*) \leq 26001 \mathit{err}^* \frac{d}{n} \tag{491}$$

Proof. Set $\delta = \text{err}^* d/n$.

Recall $\hat{M} = \|w_1\|$.

By Theorem E.12, with probability $1 - \delta$,

$$\frac{M}{2} \leq \|w_{\text{random}}\| \leq 3M/2 \tag{492}$$

$$\frac{M}{3} \leq \hat{M} \leq M \tag{493}$$

and

$$\angle(w_{\text{random}}, w^*)^2 \leq 400 \frac{\sqrt{d-1}}{Mr} \frac{d}{n} \ln(4d/\delta) \tag{494}$$

$$\leq \left(\frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3Mr} \right) \right)^2 \tag{495}$$

so

$$\angle(w_{\text{random}}, w^*) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3Mr} \right) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3\hat{M}r} \right) \tag{496}$$

Next, note that $\|w_t\| = \hat{M}$ for any t, since $\|w_1\| = \hat{M}$ and each stochastic gradient $g_t = g(w_t)$ is orthogonal to w_t and then the iterate is projected back onto $\mathcal{W}$.

Thus, define the reachable set $\mathcal{R} = \{w \in \mathcal{W} : \|w\| = \hat{M}\}$ which is the outer boundary of $\mathcal{W}$. Note $\mathcal{R}$ is not convex, but always contains the iterates w_t.

Note that $\mathbb{E}[\|g(w)\|^2] = \frac{r^2}{4}$. Set $G = \frac{r}{2}$, then the expected squared norm of the gradient is bounded by G^2.

Define $\overline{w} = \frac{\hat{M}}{M} w^*$. Then, since $\angle(w_{\text{random}}, w^*) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3\hat{M}r} \right)$, $\overline{w} \in \mathcal{R}$

For any $w \in \mathcal{R}$, $\angle(w, w_{\text{random}}) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3\hat{M}r} \right) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{Mr} \right)$. Because $\angle(w_{\text{random}}, w^*) \leq \frac{1}{2} \min\left(\frac{\pi}{2}, \frac{2}{3Mr} \right)$, $\angle(w, w^*) \leq \min\left(\frac{\pi}{2}, \frac{2}{Mr} \right)$.

Therefore, by Lemma F.1, for any $w \in \mathcal{R}$

$$\mathbb{E}[g(w)] \cdot \left(w - \frac{\|w\|}{M} w^* \right) \geq \frac{1}{6} \frac{M}{\hat{M}} \frac{r^2}{d-1} \left\| w - \frac{\|w\|}{M} w^* \right\|^2 \tag{497}$$

$$\mathbb{E}[g(w)] \cdot (w - \overline{w}) \geq \frac{1}{6} \frac{r^2}{d-1} \|w - \overline{w}\|^2 \tag{498}$$

Therefore, with $\lambda = \frac{1}{6} \frac{r^2}{d-1}$, by Lemma F.2, for $t \geq 3$,

$$\mathbb{E}\left[\|w_t - \overline{w}\|^2\right] \leq \frac{G^2}{\lambda^2 t} \tag{499}$$
$$= \frac{r^2}{4}\frac{36(d-1)^2}{r^4}\frac{1}{t} \tag{500}$$
$$= 9\frac{(d-1)^2}{r^2 t} \tag{501}$$

Because $\hat{w} = w_{n/2+1}$, for $n \geq 4$,

$$\mathbb{E}\left[\|\hat{w} - \overline{w}\|^2\right] \leq 18\frac{(d-1)^2}{r^2 n} \tag{502}$$

From Lemma 4.3 and Lemma F.3, and noting $\overline{w}$ and w^* have the same direction,

$$\mathbb{E}[\text{err}(\hat{w})] - \text{err}(w^*) \leq \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\mathbb{E}\left[\angle(\hat{w}, w^*)^2\right] \tag{503}$$
$$\leq \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\frac{(2\pi)^2}{\hat{M}^2}\mathbb{E}\left[\|\hat{w} - \overline{w}\|^2\right] \tag{504}$$
$$\leq \frac{1}{5}\frac{Mr}{\sqrt{d-1}}\frac{(2\pi)^2}{(M/3)^2}18\frac{(d-1)^2}{r^2 n} \tag{505}$$
$$\leq 1300\frac{\sqrt{d-1}}{Mr}\frac{d-1}{d}\frac{d}{n} \tag{506}$$
$$\leq 1300 \cdot 20\text{err}^*\frac{d}{n} \tag{507}$$
$$= 26000\text{err}^*\frac{d}{n} \tag{508}$$

However, this argument was predicated on an event that occurs with probability $1 - \delta$. Noting the excess error is bounded by 1, we find,

$$\mathbb{E}[\text{err}(\hat{w})] - \text{err}(w^*) \leq (1-\delta)26000\text{err}^*\frac{d}{n} + \delta \cdot 1 \tag{509}$$
$$\leq 26000\text{err}^*\frac{d}{n} + \delta \tag{510}$$
$$\leq 26001\text{err}^*\frac{d}{n} \tag{511}$$

□