Early-stopped neural networks are consistent

Ziwei Ji **Justin D. Li** **Matus Telgarsky**
<{ziweiji2,jdli3,mjt}@illinois.edu>
University of Illinois, Urbana-Champaign

Abstract

This work studies the behavior of shallow ReLU networks trained with the logistic loss via gradient descent on binary classification data where the underlying data distribution is general, and the (optimal) Bayes risk is not necessarily zero. In this setting, it is shown that gradient descent with early stopping achieves population risk arbitrarily close to optimal in terms of not just logistic and misclassification losses, but also in terms of calibration, meaning the sigmoid mapping of its outputs approximates the true underlying conditional distribution arbitrarily finely. Moreover, the necessary iteration, sample, and architectural complexities of this analysis all scale naturally with a certain complexity measure of the true conditional model. Lastly, while it is not shown that early stopping is necessary, it is shown that any univariate classifier satisfying a *local interpolation property* is inconsistent.

1 Overview and main result

Deep networks trained with gradient descent seem to have no trouble adapting to arbitrary prediction problems, and are steadily displacing stalwart methods across many domains. In this work, we provide a mathematical basis for this good performance on arbitrary binary classification problems, considering the simplest possible networks: shallow ReLU networks where only the inner (input-facing) weights are trained via vanilla gradient descent with a constant step size. The central contributions are as follows.

1. **Fully general classification tasks.** The joint distribution generating the (x, y) pairs only requires x to be bounded, and is otherwise arbitrary. In particular, the underlying distribution may be noisy, meaning the true conditional model of the labels, $\Pr[Y = 1|X = x]$, is arbitrary.

   In this setting, we show that as data, width, and training time increase, the logistic loss *measured over the population* converges to optimality over all measurable functions, which moreover implies that the induced conditional model (defined by a sigmoid mapping) converges to the true model, and the population misclassification rate also converges to optimality. This is in contrast with prior analyses of gradient descent, which either only consider the training risk [Allen-Zhu et al., 2018b, Du et al., 2019, Zou et al., 2018, Oymak and Soltanolkotabi, 2019, Song and Yang, 2019], or can only handle restricted conditional models [Allen-Zhu et al., 2018a, Arora et al., 2019, Cao and Gu, 2019, Nitanda and Suzuki, 2019, Ji and Telgarsky, 2020b, Chen et al., 2021].
2. **Adaptivity to data simplicity.** The required number of data samples, network nodes, and gradient descent iterations all shrink if the *distribution* satisfies a natural notion of simplicity: the true conditional model $\Pr[Y = 1|X = x]$ is approximated well by a low-complexity infinite-width random feature model.

Rounding out the story and contributions, firstly we present a brief toy univariate model hinting towards the necessity of early stopping: concretely, any univariate predictor satisfying a *local interpolation property* can not achieve optimal test error for noisy distributions. Secondly, our analysis is backed by a number of lemmas that could be useful elsewhere; amongst these are a

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

multiplicative error property of the logistic loss, and separately a technique to control the effects of large network width over not just a finite sample, but over the entire sphere.

1.1 Main result: optimal test error via gradient descent

The goal in this work is to minimize the logistic risk over the population: letting μ denote an arbitrary Borel measure over (x, y) pairs with compactly-supported marginal μ_x and conditional p_y, with a data sample $((x_i, y_i))_{i=1}^n$, and a function f, define the logistic loss, empirical logistic risk, and logistic risk respectively as

$$\ell(r) := \ln(1 + e^{-r}), \qquad \widehat{\mathcal{R}}(f) := \frac{1}{n}\sum_{k=1}^n \ell(y_k f(x_k)), \qquad \mathcal{R}(f) := \mathbb{E}_{x,y}\ell(yf(x)).$$

We use the logistic loss not only due to its practical prevalence, but also due to an interesting multiplicative error property which strengthens our main results (cf. Lemma B.1 and Theorem 1.1), all while being Lipschitz.

We seek to make the risk $\mathcal{R}(f)$ as small as possible: formally, we compare against the Bayes risk

$$\overline{\mathcal{R}} := \inf\left\{\mathcal{R}(f) \;:\; \text{measurable } f : \mathbb{R}^d \to \mathbb{R}\right\}.$$

While competing with $\overline{\mathcal{R}}$ may seem a strenuous goal, in fact it simplifies many aspects of the learning task. Firstly, due to the universal approximation properties of neural networks [Funahashi, 1989, Hornik et al., 1989, Cybenko, 1989, Barron, 1993], we are effectively working over the space of all measurable functions already. Secondly, as will be highlighted in the main result below, via the theory of classification calibration [Zhang, 2004, Bartlett et al., 2006], competing with the Bayes (convex) risk also recovers the true conditional model, and minimizes the misclassification loss; this stands in contrast with the ostensibly more modest goal of minimizing misclassification over a restricted class of predictors, namely the *agnostic learning* setting, which suffers a variety of computational and statistical obstructions [Goel et al., 2020a,b, Yehudai and Shamir, 2020, Frei et al., 2020].

Our predictors are shallow ReLU networks, trained via gradient descent — the simplest architecture which is not convex in its parameters, but satisfies universal approximation. In detail, letting $(a_j)_{j=1}^m$ be uniformly random ± 1 signs, $(w_j)_{j=1}^m$ with $w_j \in \mathbb{R}^d$ be standard Gaussians, and $\rho > 0$ be a *temperature*, we predict on an input $x \in \mathbb{R}^d$ with

$$f(x; \rho, a, W) := f(x; W) := \frac{\rho}{\sqrt{m}}\sum_{j=1}^m a_j \sigma_{\mathrm{r}}(w_j^{\mathsf{T}} x),$$

where $\sigma_{\mathrm{r}}(z) := \max\{0, z\}$ is the ReLU; since only W is trained, both ρ and a are often dropped. To train, we perform gradient descent with a constant step size on the empirical risk:

$$W_{i+1} := W_i - \eta \nabla \widehat{\mathcal{R}}(W_i), \qquad \text{where } \widehat{\mathcal{R}}(W) := \widehat{\mathcal{R}}\left(x \mapsto f(x; W)\right).$$

Our guarantees are for an iterate with small empirical risk and small norm: $W_{\leq t} := \arg\min\{\widehat{\mathcal{R}}(W_i) : i \leq t, \|W_i - W_0\| \leq R_{\mathrm{gd}}\}$, where R_{gd} is our *early stopping radius*: if R_{gd} is guessed correctly, our rates improve, but our analysis also handles the case $R_{\mathrm{gd}} = \infty$ where no guess is made, and indeed this is used in our final consistency analysis (a pessimistic, fully general setting).

Our goal is to show that this iterate $W_{\leq t}$ has approximately optimal *population* risk: $\mathcal{R}(W_{\leq t}) \approx \overline{\mathcal{R}}$. Certain prediction problems may seem simpler than others, and we want our analysis to reflect this while abstracting away as many coincidences of the training process as possible. Concretely, we measure simplicity via the performance and complexity of an infinite-width random feature model over the true distribution, primarily based on the following considerations.

- By measuring performance over the population, random effects of the training sample are removed, and it is impossible for the random feature model to simply revert to memorizing data, as it never sees that training data.
- The random feature model has infinite width, and via sampling can be used as a benchmark for all possible widths simultaneously, but is itself freed from coincidences of random weights.

In detail, our infinite-width random feature model is as follows. Let $\overline{U}_\infty : \mathbb{R}^d \to \mathbb{R}^d$ be an (uncountable) collection of weights (indexed by $\mathbb{R}^d$), and define a prediction mapping via

$$f(x; \overline{U}_\infty) := \int \left\langle \overline{U}_\infty(v), x\mathbb{1}[v^\intercal x \geq 0] \right\rangle \mathrm{d}\mathcal{N}(v), \qquad \text{whereby } \mathcal{R}(\overline{U}_\infty) := \mathcal{R}(x \mapsto f(x; \overline{U}_\infty)).$$

Note that for each Gaussian random vector $v \sim \mathcal{N}$, we construct a *random feature* $x \mapsto x\mathbb{1}[v^\intercal x \geq 0]$. This particular choice is simply the gradient of a corresponding ReLU $\nabla_v \sigma_{\mathrm{r}}(v^\intercal x)$, and is motivated by the NTK literature [Jacot et al., 2018, Li and Liang, 2018, Du et al., 2019]. A similar object has appeared before in NTK convergence analyses [Nitanda and Suzuki, 2019, Ji and Telgarsky, 2020b], but the conditions on $\overline{U}_\infty$ were always strong (e.g., data separation with a margin).

What, then, does it mean for the data to be simple? In this work, it is when there exists a $\overline{U}_\infty$ with $\mathcal{R}(\overline{U}_\infty) \approx \overline{\mathcal{R}}$, and moreover $\overline{U}_\infty$ has low norm; for technical convenience, we measure the norm as the maximum over individual weight norms, meaning $\sup_v \|\overline{U}_\infty(v)\|$. To measure approximability, for sake of interpretation, we use the *binary Kullback-Leibler divergence (KL)*: defining a conditional probability model ϕ_∞ corresponding to $\overline{U}_\infty$ via

$$\phi_\infty(x) := \phi(f(x; \overline{U}_\infty)), \qquad \text{where } \phi(r) := \frac{1}{1 + \exp(-r)},$$

then the binary KL can be written as

$$\mathcal{K}_{\mathsf{bin}}(p_y, \phi_\infty) := \int \left(p_y \ln \frac{p_y}{\phi_\infty} + (1 - p_y) \ln \frac{1 - p_y}{1 - \phi_\infty} \right) \mathrm{d}\mu_x = \mathcal{R}(\overline{U}_\infty) - \overline{\mathcal{R}}.$$

This relationship between binary KL and the excess risk is a convenient property of the logistic loss, which immediately implies calibration as a consequence of achieving the optimal risk.

The pieces are all in place to state our main result.

Theorem 1.1. *Let width $m \geq \ln(emd)$, temperature $\rho > 0$, and reference model $\overline{U}_\infty$ be given with $R := \max\{4, \rho, \sup_v \|\overline{U}_\infty(v)\|\} < \infty$, and define a corresponding conditional model $\phi_\infty(x) := \phi(f(x; \overline{U}_\infty))$. Let optimization accuracy ϵ_{gd} and radius $R_{\mathrm{gd}} \geq R/\rho$ be given, define effective radius $B := \min\left\{R_{\mathrm{gd}}, \ \frac{3R}{\rho} + \frac{4e}{\rho}\sqrt{t}\sqrt{e^{\tau_0}\mathcal{R}(\overline{U}_\infty) + R\tau_n}\right\}$, and generalization, linearization, and sampling errors (τ_n, τ_1, τ_0) as*

$$\tau_n := \widetilde{\mathcal{O}}\left(\frac{(d\ln(1/\delta))^{3/2}}{\sqrt{n}}\right), \ \tau_1 := \widetilde{\mathcal{O}}\left(\frac{\rho B^{4/3}\sqrt{d\ln(1/\delta)}}{m^{1/6}}\right), \ \tau_0 := \widetilde{\mathcal{O}}\left(\rho\ln(1/\delta) + \frac{\sqrt{d\ln(1/\delta)}}{m^{1/4}}\right),$$

where it is assumed $\tau_1 \leq 2$, and $\widetilde{\mathcal{O}}$ hides constants and $\ln(nmd)$. Choose step size $\eta := 4/\rho^2$, and run gradient descent for $t := 1/(8\epsilon_{\mathrm{gd}})$ iterations, selecting iterate $W_{\leq t} := \arg\min\{\widehat{\mathcal{R}}(W_i) : i \leq t, \|W_i - W_0\| \leq R_{\mathrm{gd}}\}$. Then, with probability at least $1 - 25\delta$,

$$\begin{aligned}
& \mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} && \textit{(logistic error)} \\
\leq \ & \mathcal{K}_{\mathsf{bin}}(p_y, \phi_\infty) + \left(e^{\tau_1 + \tau_0} - 1\right)\mathcal{R}(\overline{U}_\infty) && \textit{(reference model error)} \\
+ \ & e^{\tau_1} R^2 \epsilon_{\mathrm{gd}} && \textit{(optimization error)} \\
+ \ & e^{\tau_1}(\rho B + R)\tau_n && \textit{(generalization error)},
\end{aligned}$$

where the classification and calibration errors satisfy

$$\begin{aligned}
& \mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} && \textit{(logistic error)} \\
\geq \ & 2\int \left(\phi(f(x; W_{\leq t})) - p_y\right)^2 \mathrm{d}\mu_x(x) && \textit{(calibration error)} \\
\geq \ & \frac{1}{2}\left(\mathcal{R}_{\mathrm{z}}(W_{\leq t}) - \overline{\mathcal{R}}_{\mathrm{z}}\right)^2 && \textit{(classification error)}.
\end{aligned}$$

Lastly, for any $\epsilon > 0$, there exists $\overline{U}_\infty^{(\epsilon)}$ with $\sup_v \|\overline{U}_\infty^{(\epsilon)}(v)\| < \infty$ and whose conditional model $\phi_\infty^{(\epsilon)}(x) := \phi(f((x,1)/\sqrt{2}; \overline{U}_\infty^{(\epsilon)}))$ satisfies $\mathcal{K}_{\mathsf{bin}}(p_y, \phi_\infty^{(\epsilon)}) \leq \epsilon$.

Remark 1.1. The key properties of Theorem 1.1 are as follows.

1. **(Achieving error $\mathcal{O}(\epsilon)$ in three different regimes.)** As Theorem 1.1 is quite complicated, consider three different situations, which vary the reference model $\overline{U}_\infty$ and its norm upper bound $R := \max\{4, \rho, \sup_v \|\overline{U}_\infty(v)\|\} < \infty$, as well as the early stopping radius R_{gd}. Let target population (excess) risk $\epsilon > 0$ be given, set $\epsilon_{\text{gd}} = \epsilon$ and $t = 1/(8\epsilon_{\text{gd}})$ as in Theorem 1.1, and suppose $n \geq 1/\epsilon^2$ samples: in each of the three following settings, the other parameters parameters (namely ρ and m) will be chosen to ensure a final error $\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} = \mathcal{O}(\epsilon)$.

   (a) **(Easy data.)** Suppose a setting with *easy data*: specifically, suppose that for chosen target accuracy $\epsilon > 0$, there exists $\overline{U}_\infty$ with $\mathcal{K}_{\text{bin}}(p_y, \phi_\infty) = \mathcal{R}(\overline{U}_\infty) - \overline{\mathcal{R}} \leq \mathcal{R}(\overline{U}_\infty) \leq \epsilon$. If we set $\rho = 1$ and $m \geq R^8$, then (τ_n, τ_1, τ_0) are all constant, and we get a final bound $\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} = \mathcal{O}(\epsilon)$.
   Note crucially that $m \approx R^8$ sufficed for this setting; this was a goal of the present analysis, as it recovers the *polylogarithmic width* analyses from prior work [Ji and Telgarsky, 2020b, Chen et al., 2021]. Those works however either used a separation condition due to Nitanda and Suzuki [2019] in the shallow case, or an assumption on the approximation properties of the sampled weights (a random variable) in the deep case, and thus the present analysis provides not just a re-proof, but a simplification and generalization. This was the motivation for the strange *multiplicative* form of the errors in Theorem 1.1: had we used the more common additive errors with standard linearization tools, a polylogarithmic width proof would fail.

   (b) **(General data, clairvoyant early stopping radius R_{gd}.)** Suppose that we are in the general noisy case, meaning any $\overline{U}_\infty$ we pick has a large error $\mathcal{K}_{\text{bin}}(p_y, \phi_\infty)$, but we magically know the R corresponding to a good $\overline{U}_\infty$, and can choose $R_{\text{gd}} = R/\rho$. Unlike the previous case, to achieve some target error ϵ, we need to work harder to control the term $\left[\exp(\tau_1 + \tau_0) - 1\right] \mathcal{R}(\overline{U}_\infty)$, since we no longer have small $\mathcal{R}(\overline{U}_\infty)$; to this end, since $\tau_1 = \widetilde{\mathcal{O}}(R^{4/3}/(m\rho^2)^{1/6})$ and $\tau_0 = \widetilde{\mathcal{O}}(\rho + 1/m^{1/4})$, choosing $\rho = m^{-1/8}$ and $m = 1/\epsilon^8$ gives $\tau_1 = \widetilde{\mathcal{O}}(\epsilon)$ and $\tau_0 = \widetilde{\mathcal{O}}(\epsilon)$, and together $\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} = \mathcal{O}(\epsilon)$.

   (c) **(General data, worst-case early stopping.)** Suppose again the case of general noisy data with large error $\mathcal{K}_{\text{bin}}(p_y, \phi_\infty)$ for any $\overline{U}_\infty$ we pick, but now suppose we have no early stopping hint, and pessimistically set $R_{\text{gd}} = \infty$. As a consequence of all of this, the term B can scale as $t^{2/3}/\rho = 1/(\rho\epsilon^{2/3})$, thus to control $\tau_1 = \widetilde{\mathcal{O}}((1/\epsilon)^{2/3}/(m\rho^2)^{1/6})$ and $\tau_0 = \widetilde{\mathcal{O}}(\rho + 1/m^{1/4})$, we can again choose $\rho = m^{-1/8}$, but need a larger width $m = 1/\epsilon^{40/3}$. Together, we once again achieve population excess risk $\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} = \mathcal{O}(\epsilon)$.

   Summarizing, a first key point is that arbitrarily small excess risk $\mathcal{O}(\epsilon)$ is always possible; as discussed, this is in contrast to prior work, which either only gave training error guarantees, or required restrictive conditions for small test error. A second key point is that the parameters of the bound, most notably the required width, will shrink greatly when either the data is easy, or an optimal stopping radius R_{gd} is known.

2. **(Consistency.)** *Consistency* is a classical statistical goal of achieving the optimal test error almost surely *over all possible predictors* as $n \to \infty$; here it is proved as a consequence of Theorem 1.1, namely the preceding argument that we can achieve excess risk $\mathcal{O}(\epsilon)$ even with general prediction problems and no early stopping hints ($R_{\text{gd}} = \infty$). The consistency guarantee is stated formally in Corollary 2.3. The statement takes the width to infinity, and demonstrates another advantage of using an infinite-width reference model: within the proof, after fixing a target accuracy, the reference model is *fixed* and used for all widths simultaneously.

3. **(Non-vacuous generalization, and an estimate of R.)** There is extensive concern throughout the community that generalization estimates are hopelessly loose [Neyshabur et al., 2014, Zhang et al., 2016, Dziugaite and Roy, 2017]; to reduce the concern here, we raise two points. Firstly, these concerns usually involve explicit calculations of generalization bounds which have terms scaling with some combination of $\|W\|$ (not $\|W - W_0\|$) and m; e.g,. one standard bound has spectral norms $\|W\|_2$ and $(2,1)$ matrix norms $\|(W - W_0)^\intercal\|_{2,1}$, which are upper bounded by $\|W - W_0\|\sqrt{m}$ [Bartlett et al., 2017]. By contrast, the present work uses a new generalization

bound technique (cf. Lemma B.8) which first *de-linearizes* the network, then applies a *linear generalization bound* which has only $\|W - W_0\|$ and no explicit $\mathrm{poly}(m)$, and then *re-linearizes*.

Secondly, there may still be concern that the story here is broken due to the term R, and namely the non-existence of good choices for $\overline{U}_\infty$. For this, we conducted a simple experiment. Noting that we can freeze the initial features and train linear predictors of the form $f^{(0)}(x; V)$ for weights $V \in \mathbb{R}^{m \times d}$ (cf. section 1.4), and that the performance converges to the infinite-width performance as $m \to \infty$, we fixed a large width and trained two prediction tasks: an *easy* task of MNIST 1 vs 5 until $R_{\text{easy}}/\sqrt{n} \approx 1/2$, and a *hard* task of MNIST 3 vs 5 until $R_{\text{hard}}/\sqrt{n} \approx 1/2$. After training, we obtained test error $\mathcal{R}(V_{\text{easy}}) \approx 0.01$ and $\mathcal{R}(V_{\text{hard}}) \approx 0.08$. Plugging all of these terms back in to the bound, firstly these techniques can yield a non-vacuous generalization bound, secondly they do not exhibit bad scaling with large width, and thirdly they do reflect the difficulty of the problem, as desired.

4. **(Early stopping and the NTK.)** As discussed above, when the data is noisy, the method is explicitly early stopped, either by clairvoyantly choosing R_{gd}, or by making t small. In this setting, the optimization accuracy ϵ_{gd} is an *excess* empirical risk, meaning in particular that 0 training error (the *interpolation regime* [Belkin et al., 2018a]) will *not* be reached. This is in stark contrast to standard NTK analyses [Allen-Zhu et al., 2018b], which guarantee zero training error, but can not ensure good test error in general. Since the NTK itself is an early stopping (as in, if one continues to optimizes, one exits the NTK), then the early stopping in this work is even earlier than the NTK early stopping; this situation is summarized in Figure 1 in the appendix, and will be revisited for the lower bound in Section 1.2.

5. **(Classification and calibration.)** The relationship to classification and calibration errors is merely a restatement of existing results [Zhang, 2004, Bartlett et al., 2006], though it is reproved here in an elementary way for the special case of the logistic loss. Similarly, the guarantee that $\mathcal{K}_{\text{bin}}(p_y, \phi_\infty^{(\epsilon)})$ can be made arbitrarily small is also not a primary contribution, and indeed most of the heavy lifting is provided both by prior work in neural network approximation [Barron, 1993], and by the existing and reliable machinery for proving consistency [Schapire and Freund, 2012]. As such, the consistency result is stated only much later in Corollary 2.3, and our focus is on the exact risk guarantees in Theorem 1.1.

6. **(Inputs with bias: $(x, 1)/\sqrt{2} \in \mathbb{R}^{d+1}$.)** The end of Theorem 1.1 appends a constant to the input (and rescales), which simulates a bias term inside each ReLU; this is necessary since our models are (sigmoid mappings of) homogeneous functions, whereas p_y is general. Biases are also simulated in this way in the consistency result in Corollary 2.3.

Further discussion of Theorem 1.1, including the formal consistency result (cf. Corollary 2.3) and a proof sketch, all appear in Section 2. Full proofs appear in the appendices.

1.2 Should we early stop?

Theorem 1.1 uses early stopping: it can blow up if $\overline{\mathcal{R}} > 0$ and the two gradient descent parameters R_{gd} and $1/\epsilon_{\text{gd}}$ are taken to ∞ in an uncoordinated fashion. Part of this is purely technical: as with many neural network optimization proofs, the analysis breaks when far from initialization. It is of course natural to wonder what happens if one trains indefinitely, entering the actively-studied *interpolation regime* [Belkin et al., 2018b,a, Bartlett et al., 2019]. Furthermore, there is evidence that gradient descent on shallow networks limits towards a particular interpolating choice, one with large margins [Soudry et al., 2018, Lyu and Li, 2020, Chizat and Bach, 2020, Ji and Telgarsky, 2020a]. Is this behavior favorable?

While we do not rule out that the interpolating solutions found by neural networks perform well, we show that at least in the low-dimensional (univariate!) setting, if a prediction rule perfectly labels the data and is not too wild between training points, then it is guaranteed to achieve poor test loss on noisy problems. This negative observation is not completely at odds with the interpolation literature, where the performance of some rules improves with dimension [Belkin et al., 2018b].

Proposition 1.2. *Given a finite sample $((x_i, y_i))_{i=1}^n$ with $x_i \in \mathbb{R}$ and $y_i \in \{\pm 1\}$, let $\mathcal{F}_n$ denote the collection of* local interpolation rules *(cf. Figure 2 in the appendix): letting $x_{(i)}$ index examples in*

sorted order, meaning $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$*, define* $\mathcal{F}_n$ *as*

$$\mathcal{F}_n := \big\{ f : \mathbb{R} \to \mathbb{R} \ : \ \forall i \ f(x_{(i)}) = y_{(i)}, \text{ and} \\ \text{if } y_{(i)} = y_{(i+1)}, \text{ then } \inf_{\alpha \in [0,1]} f\left(\alpha x_{(i)} + (1-\alpha) x_{(i+1)}\right) y_{(i)} > 0 \big\}.$$

Then there exists a constant $c > 0$ *so that with probability at least* $1 - \delta$ *over the draw of* $((x_i, y_i))_{i=1}^n$ *with* $n \geq \ln(1/\delta)/c$*, every* $f \in \mathcal{F}_n$ *satisfies* $\mathcal{R}_z(f) \geq \bar{\mathcal{R}}_z(f) + c$*.*

Although a minor contribution, this result will be discussed briefly in Section 3, with detailed proofs appearing in the appendices. For a similar discussion for nearest neighbor classifiers albeit under a few additional assumptions, see [Nakkiran and Bansal, 2021].

1.3 Related work

Analyses of gradient descent. The proof here shares the most elements with recent works whose width could be polylogarithmic in the sample size and desired target accuracy $1/\epsilon$ [Ji and Telgarsky, 2020b, Chen et al., 2021]. Similarities include using a regret inequality as the core of the proof, using an infinite-width target network [Nitanda and Suzuki, 2019, Ji and Telgarsky, 2020b], and using a *linearization inequality* [Chen et al., 2021, Allen-Zhu et al., 2018b]. On the technical side, the present work differs in the detailed treatment of the logistic loss, and in the linearization inequality which is extended to hold over the population risk; otherwise, the core gradient descent analysis here is arguably simplified relative to these prior works. It should be noted that the use of a regret inequality here and in the previous works crucially makes use of a negated term which was dropped in some classical treatments; this trick is now re-appearing in many places [Orabona and Pál, 2021, Frei et al., 2020].

There are many other, somewhat less similar works in the vast literature of gradient descent on neural networks, in particular in the neural tangent regime [Jacot et al., 2018, Li and Liang, 2018, Du et al., 2019]. These works often handle not only training error, but also testing error [Allen-Zhu et al., 2018a, Arora et al., 2019, Cao and Gu, 2019, Nitanda and Suzuki, 2019, Ji and Telgarsky, 2020b, Chen et al., 2021]. As was mentioned before, these works do not appear to handle arbitrary target models; see for instance the modeling discussion in [Arora et al., 2019, Section 6]. As another interesting recent example, some works explicitly handle certain noisy conditional models, but with error terms that do not go to zero in general [Liang et al., 2021].

Consistency. Consistency of deep networks with classification loss and *some* training procedure is classical; e.g., in [Farago and Lugosi, 1993], the authors show that it suffices to run a computationally intractable algorithm on an architecture chosen to balance VC dimension and universal approximation. Similarly, the work here makes use of Barron's superposition analysis in an infinite-width form to meet the Bayes risk [Barron, 1993, Ji et al., 2020b]. The statistics literature has many other works giving beautiful analyses of neural networks, e.g., even with minimax rates [Schmidt-Hieber, 2017], though it appears this literature generally does not consider gradient descent and arbitrary classification objectives.

In the boosting literature, most consistency proofs only consider classification loss [Bartlett and Traskin, 2007, Schapire and Freund, 2012], though there is a notable exception which controls the convex loss (and thus calibration), although the algorithm has a number of modifications [Zhang and Yu, 2005]. In all these works, arbitrary p_y are not handled explicitly as here, but rather *implicitly* via assumptions on the expressiveness of the weak learners. One exception is the logistic loss boosting proof of Telgarsky [2013], which explicitly handles measurable p_y via Lusin's theorem as is done here, but ultimately the proof only controls classification loss.

Following the arXiv posting of this work, a few closely related works appeared. Firstly, Richards and Kuzborskij [2021] show that the expected excess risk can scale with $\|W_t - W_0\|_{\mathrm{F}}/n^\alpha$, though in contrast with the present work, it is not shown that this ratio can go to zero for arbitrary prediction problems, and moreover the bound is in expectation only. Secondly, the work of Braun et al. [2021] is even closer, however it requires a condition on the Fourier spectrum of the conditional model p_y, which is circumvented here via a more careful Fourier analysis due to Ji et al. [2020b].

Calibration. There is an increasing body of work considering the (in)ability of networks trained with the logistic loss to recover the underlying conditional model. Both on the empirical side [Guo

et al., 2017] and on the theoretical side [Bai et al., 2021], the evidence is on the side of the logistic loss doing poorly, specifically being *overconfident*, meaning the sigmoid outputs are too close to 0 or 1. This overconfident regime corresponds to large margins; indeed, since gradient descent can be proved in some settings to exhibit unboundedly large unnormalized margins on all training points [Lyu and Li, 2020], the sigmoid mapping of the predictions will necessarily limit to exactly 0 or 1. On the other hand, as mentioned in [Bai et al., 2021], regularization suffices to circumvent this issue. In the present work, a combination of early stopping and small temperature are employed. As mentioned before, calibration is proved here as an immediate corollary of meeting the optimal logistic risk via classification calibration [Zhang, 2004, Bartlett et al., 2006].

1.4 Further notation and technical background

The loss ℓ, risks $\mathcal{R}$ and $\widehat{\mathcal{R}}$, and network f have been defined. The misclassification risk $\mathcal{R}_z(f) = \Pr[\operatorname{sgn}(f(X)) \neq Y]$ appeared in Theorem 1.1, where $\operatorname{sgn}(f(x)) = 2 \cdot \mathbb{1}[f(x) \geq 0] - 1$.

Next, consider the "gradient" of f with respect to weights W:

$$\nabla f(x; W) := \frac{\rho}{\sqrt{m}} \sum_{j=1}^{m} a_j \mathbb{1}[w_j^\intercal x \geq 0] \boldsymbol{e}_j x^\intercal;$$

it may seem the nondifferentiability at 0 is concerning, but in analyses close to initialization (as is the one here), few activations change, and their behavior is treated in a worst-case fashion. Note that, as is easily checked with this expression, $\|\nabla f(W)\| \leq \rho$, which is convenient in many places in the proofs. Here $\|\cdot\|$ denotes the Frobenius norm; $\|\cdot\|_2$ will denote the spectral norm.

Given weight matrix W_i at time i, let $(w_{i,j}^\intercal)_{j=1}^m$ refer to its rows. Define features $f^{(i)}$ at time i and a corresponding empirical risk $\widehat{\mathcal{R}}^{(i)}$ using the features at time i as

$$\begin{aligned} f^{(i)}(x; V) &:= \left\langle \nabla f(x; W_i), V \right\rangle = \frac{\rho}{\sqrt{m}} \sum_j a_j v_j^\intercal x \mathbb{1}[w_{i,j}^\intercal x \geq 0], \\ \widehat{\mathcal{R}}^{(i)}(x; V) &:= \widehat{\mathcal{R}}(x \mapsto f^{(i)}(x; V)). \end{aligned}$$

By 1-homogeneity of the ReLU, $f^{(i)}(x; W_i) = f(x; W_i)$, which will also be used often. These features at time i, meaning $f^{(i)}$ and $\widehat{\mathcal{R}}^{(i)}$, are very useful in analyses near initialization, as they do not change much. As such, $f^{(0)}$ and $\mathcal{R}^{(0)}$ and $\widehat{\mathcal{R}}^{(0)}$ will all appear often as well.

To be a bit pedantic about the measure μ: as before, there is a joint distribution μ, which is over the Borel σ-algebra on $\mathbb{R}^d \times \{\pm 1\}$, where $\|x\| \leq 1$ almost surely. This condition suffices to grant both a *disintegration* of μ into marginal μ_x and conditional p_y [Kallenberg, 2002, Chapter 6], and also Lusin's theorem [Folland, 1999, Theorem 7.10], which is used to switch from a measurable function to a continuous one in the consistency proof (cf. Corollary 2.3).

2 Discussion and proof sketch of Theorem 1.1

This section breaks down the proof and discussion into four subsections: a section with common technical tools, then sections for the analysis of generalization, optimization, and approximation.

2.1 Key technical lemmas

There are two main new technical ideas which power many parts of the proofs: a multiplicative error property of the logistic loss, and a *linearization over the sphere*.

The logistic loss property is simple enough: for any $a \geq b$, it holds that $\ell(-a)/\ell(-b) \leq \exp(a - b)$. On the surface, this seems innocuous, but this simple inequality allows us to reprove existing polylogarithmic width results for easy data [Ji and Telgarsky, 2020b, Chen et al., 2021], however making use of a proof scheme which is slightly more standard, or at the very least more apparently a smooth convex proof with just this one special property of the logistic loss (as opposed to a few special properties).

The second tool is more technical, and is used crucially in many places in the proof. Many prior analyses near initialization bound the quantity

$$f(x;V) - f(x;W) - \left\langle \nabla f(x;W), V - W \right\rangle,$$

where V and W are both close to initialization [Allen-Zhu et al., 2018b, Cao and Gu, 2019, Chen et al., 2021]. These proofs are typically performed on a fixed example x_k, and then a union bound carries them over to the whole training set. Here, instead, such a bound is extended to hold *over the entire sphere*, as follows.

Lemma 2.1 (Simplification of Lemma B.7)**.** *Let scalars $\delta > 0$ and $R_V \geq 1$ and $R_B \geq 0$ be given.*

1. *With probability at least $1 - 3n\delta$,*

$$\sup_{\substack{\|W_i - W_0\| \leq R_V \\ \|W_j - W_0\| \leq R_V \\ \|B - W_0\| \leq R_B}} \frac{\widehat{\mathcal{R}}^{(i)}(B)}{\widehat{\mathcal{R}}^{(j)}(B)} \leq \exp\left(\frac{6\rho \left(R_B + 2R_V\right) R_V^{1/3} \ln(e/\delta)^{1/4}}{m^{1/6}} \right).$$

2. *Suppose $m \geq \ln(edm)$. With probability at least $1 - (1 + 3(d^2 m)^d)\delta$,*

$$\sup_{\|W_i - W_0\| \leq R_V} \frac{\mathcal{R}(W_i)}{\mathcal{R}^{(0)}(W_i)} \leq \exp\left(\frac{25\rho R_V^{4/3} \sqrt{\ln(edm/\delta)}}{m^{1/6}} \right).$$

The preceding lemma combines both the linearization technique and the multiplicative error property: it bounds how much the empirical and true risk change for a fix weight matrix if we swap in and out the features at different iterations. That these bounds are a ratio is due to the multiplicative error property. That the second part holds over the true risk, in particular controlling behavior over all $\|x\| \leq 1$, is a consequence of the new more powerful linearization technique. This linearization over the sphere is used crucially in three separate places: we use it when controlling the range in the generalization proofs, when *de-linearizing* after generalization, and when sampling from the infinite-width model $\overline{U}_\infty$. The method of proof is inspired by the concept of *co-VC dimension* [Gurvits and Koiran, 1995]: the desired inequality is first union bounded over a cover of the sphere, and then relaxed to all points on the sphere. A key difficulty here is the non-smoothness of the ReLU, and a key lemma establishes a smoothness-like inequality (cf. Lemma B.5). These techniques appear in full in the appendices.

2.2 Generalization analysis

The generalization analysis ends up being easy thanks to the multiplicative error control in Lemma B.1, and the aforementioned linearization over all points in the sphere. Specifically, to prove generalization, the network is first linearized, then generalization *of linear predictors* is applied, and then de-linearization is applied on the population risk side. This generalization bound only pays logarithmically in the width m.

Typically the easiest step in proving generalization is to provide a worst-case estimate on the range of the predictor. Here, since there is a goal of controlling the logistic loss over the population, brute forcing this range estimate incurs a polynomial dependence on network width. The solution here is again to apply the aforementioned Lemma B.3; the generalization statement appears in full as Lemma B.8 together with its proof in the appendices.

2.3 Gradient descent analysis

A common tool in linear prediction is the regret inequality

$$\|v_t - z\|^2 + 2\eta \sum_{i<t} \widehat{\mathcal{R}}(v_{i+1}) \leq \|v_0 - z\|^2 + 2t\eta \widehat{\mathcal{R}}(z),$$

which can be derived by expanding the square in $\|v_t - z\|^2$ and applying smoothness and convexity. The term $\|v_t - z\|^2$ is often dropped, but can be used in a very convenient way: by the triangle inequality, if $\|v_t - v_0\| \geq 2\|z - v_0\|$, then the norm terms above may be canceled from both sides,

which leaves only the empirical risk terms; overall, this argument ensures both small norm and small empirical risk. This idea has appeared in a variety of works [Shamir, 2020, Ji et al., 2020a], and is used here to provide a convenient norm control, allowing linearization and all other proof parts to go through. Combining this idea with the earlier generalization analysis and a few other minor tricks gives the following bounds, which in turn provide most of Theorem 1.1.

Lemma 2.2. *Let temperature $\rho > 0$, step size $\eta \leq 4/\rho^2$, optimization accuracy $\epsilon_{\text{gd}} > 0$, radius $R_{\text{gd}} > 0$, network width $m \geq \ln(emd)$, reference matrix $Z \in \mathbb{R}^{m\times d}$, corresponding scalar $R_Z \leq R_{\text{gd}}$ where $R_Z \geq \max\{1, \eta\rho, \|W_0 - Z\|\}$, and $t \geq 1/(2\eta\rho^2\epsilon_{\text{gd}})$ be given; correspondingly define $W_{\leq t} := \arg\min\{\widehat{\mathcal{R}}(W_i) : i \leq t, \|W_i - W_0\| \leq R_{\text{gd}}\}$. Define effective radius $B := \min\left\{R_{\text{gd}},\ 3R_Z + 2e\sqrt{\eta t \widehat{\mathcal{R}}^{(0)}(Z)}\right\}$, and linearization and generalization errors*

$$\tau := \frac{25\rho B^{4/3}\sqrt{d\ln(em^2d^3/\delta)}}{m^{1/6}}, \qquad \tau_n := \frac{80\left(d\ln(em^2d^3/\delta)\right)^{3/2}}{\sqrt{n}},$$

and suppose $\tau \leq 2$. Then, with probability at least $1 - 3n\delta$, the selected iterate $W_{\leq t}$ satisfies $\|W_{\leq t} - W_0\| \leq B$, along with the empirical risk guarantee

$$\widehat{\mathcal{R}}(W_{\leq t}) \leq e^{2\tau}\widehat{\mathcal{R}}^{(0)}(Z) + e^{\tau}(\rho R_Z)^2\epsilon_{\text{gd}},$$

and by discarding an additional 16δ failure probability, then $\widehat{\mathcal{R}}^{(0)}(Z) \leq \mathcal{R}^{(0)}(Z) + \rho R_Z\tau_n$, and

$$\mathcal{R}(W_{\leq t}) \leq e^{4\tau}\mathcal{R}^{(0)}(Z) + e^{3\tau}(\rho R_Z)^2\epsilon_{\text{gd}} + e^{4\tau}(B + R_Z)\rho\tau_n.$$

This version of the statement, unlike Theorem 1.1, features an arbitrary reference *matrix* Z. This is powerful, though it can be awkward, since W_0 is a random variable.

2.4 Approximation analysis, consistency, and the proof of Theorem 1.1

Rather than trying to reason about good predictors which may happen to be close to random initialization, the approach here is instead to start from deterministic predictors over the population (e.g., $\overline{U}_\infty$), and to use their structure to construct approximants near the initial iterate, the random matrix W_0. Specifically, the approach here is fairly brute force: given initial weights W_0 with rows $(w_{0,j}^\intercal)_{j=1}^m$, the rows $(\overline{u}_j)_{j=1}^m$ of the finite width reference matrix $\overline{U} \in \mathbb{R}^{m\times d}$ intended to mimic $\overline{U}_\infty$ (which is after all a mapping $\overline{U}_\infty : \mathbb{R}^d \to \mathbb{R}^d$) are simply

$$\overline{u}_j := \frac{a_j\overline{U}_\infty(w_{0,j})}{\rho\sqrt{m}} + w_{0,j}. \tag{1}$$

By construction, $\|\overline{U} - W_0\| \leq R/\rho$, where $R := \sup_v \|\overline{U}_\infty(v)\|$. To argue that $\mathcal{R}^{(0)}(\overline{U})$ and $\mathcal{R}(\overline{U}_\infty)$ are close, the abstract control over the sphere in Lemma B.3 is again used. Plugging this $\overline{U}$ into Lemma 2.2 and introducing $\mathcal{K}_{\text{bin}}(p_y, \phi_\infty)$ via Lemma B.1 gives the first part of Theorem 1.1, and the second part of Theorem 1.1 is also from Lemma B.1.

It remains to prove that for any p_y, there exists $\overline{U}_\infty$ with $\phi(x \mapsto f((x,1)/\sqrt{2}; \overline{U}_\infty)) \approx p_y$ (we must include a bias term, as mentioned in Remark 1.1). If p_y were continuous, there is a variant of Barron's seminal universal approximation construction which explicitly gives an infinite-width network of the desired form [Barron, 1993, Ji et al., 2020b]. To address continuity is even easier: *Lusin's theorem* [Folland, 1999, Theorem 7.10] lets us take the measurable function p_y, and obtain a continuous function that agrees with it on all but a negligible fraction of the domain. This completes the proof.

As mentioned, a key property of the reference model $\overline{U}_\infty$ is that it depends on neither the random sampling of data, nor the random sampling of weights. This vastly simplifies the proof of consistency, where the proof scheme first fixes an $\epsilon > 0$ and chooses a $\overline{U}_\infty$, and leaves it fixed as m and n vary.

Corollary 2.3. *Let early stopping parameter $\xi \in (0,1)$ be given, and for each sample size n, define a weight matrix $\widehat{W}_n \in \mathbb{R}^{m^{(n)}\times(d+1)}$ and corresponding conditional probability model $\widehat{\phi}_n(x) := \phi(f((x,1)/\sqrt{2}; \widehat{W}_n))$ as follows. For each sample size n, let $(W_i^{(n)})_{i\geq 0}$ denote the corresponding sequence of gradient descent iterates obtained with parameter choices $\rho^{(n)} := (m^{(n)})^{-1/8}$, and*

$m^{(n)} := n^{\frac{40}{3}(1-\xi)}$, *and* $\eta^{(n)} := 4/(\rho^{(n)})^2$, *and* $\epsilon_{\text{gd}}^{(n)} := n^{\xi-1}$, *and* $t^{(n)} := n^{1-\xi}/8$, *and choose the empirical risk minimizer over the sequence, meaning* $\widehat{W}_n := \arg\min\left\{\widehat{\mathcal{R}}(W_i^{(n)}) : i \le t^{(n)}\right\}$ *(in the notation of Theorem 1.1, this is* $W_{\le}t$ *with* $R_{\text{gd}} = \infty$*). Then*

$$\mathcal{R}(\widehat{W}_n) \longrightarrow \overline{\mathcal{R}} \text{ a.s.}, \qquad \mathcal{R}_{\text{z}}(\widehat{W}_n) \longrightarrow \overline{\mathcal{R}}_{\text{z}} \text{ a.s.}, \qquad \widehat{\phi}_n \xrightarrow{L_2(\mu_x)} p_y \text{ a.s.},$$

where the last convergence is in the $L_2(\mu_x)$ *metric.*

The use of a parameter $\xi \in (0,1)$ is standard in similar consistency results; see for instance the analogous parameter in the consistency analysis of AdaBoost [Bartlett and Traskin, 2007]. Proofs, as usual, are in the appendices.

3 Discussion and proof sketch of Proposition 1.2

Proposition 1.2 asserts that univariate *local interpolation rules* — predictors which perfectly fit the data, and are not too wild between data points of the same label — will necessarily achieve suboptimal population risk. The proof idea seems simple enough: if the true conditional probability p_y is not one of $\{0, 1/2, 1\}$ everywhere, and is also continuous, then there must exist a region where it is well separated from these three choices. It seems natural that a constant fraction of the data in these regions will form adjacent pairs with the wrong label; a local interpolation rule will fail on exactly these adjacent noisy pairs, which suffices to give the bound. In reality, while this is indeed the proof scheme followed here, the full proof must contend with many technicalities and independence issues. It appears in the appendices.

While the motivation in Section 1.2 focused on neural networks which interpolate, and also maximum margin solutions, the behavior on this noisy univariate data is also well-illustrated by k-nearest-neighbors classifiers (k-nn). Specifically, 1-nn is a local interpolant, and Proposition 1.2 applies. On the other hand, choosing $k = \Theta(\ln(n))$ is known to provide enough smoothing to achieve consistency and avoid interpolation [Devroye et al., 1996].

It should be stressed again that even if the remaining pieces could be proved to apply this result to neural networks, namely necessitating early stopping, it would still be a univariate result only, leaving open many interesting possibilities in higher dimensions.

4 Concluding remarks and open problems

Empirical performance. Does the story here match experiments? E.g., is it often the case that if a neural network performs well, then so does a random feature model? Do neural networks fail on noisy data if care is not taken with temperature and early stopping? Most specifically, is this part of what happens in existing results reporting such failures [Guo et al., 2017]?

Temperature parameter ρ. Another interesting point of study is the temperature parameter ρ. It arises here in a fairly technical way: if p_y is often close to $1/2$, then the random initialization of W_0 gets in the way of learning p_y. The temperature ρ is in fact a brute-force method of suppressing this weight initialization noise. On the other hand, temperature parameters are common across many works which rely heavily on the detailed real-valued outputs of sigmoid and softmax mappings; e.g., in the distillation literature [Hinton et al., 2015]. The temperature also plays the same role as the scale parameter in the *lazy training* regime [Chizat and Bach, 2019]. Is ρ generally useful, and does the analysis here relate to its practical utility?

Random features, and going beyond the NTK. The analysis here early stops before the feature learning begins to occur. How do things fare outside the NTK? Is there an analog of Theorem 1.1, still stopping shy of the interpolation pitfalls of Proposition 1.2, but managing to beat random features with some generality?

The logistic loss. One reason the logistic is used here is its simple interplay with calibration (e.g., see the elementary proof of Lemma B.1, as compared with the full machinery of classification calibration [Zhang, 2004, Bartlett et al., 2006]). The other key reason was the multiplicative error property Lemma B.1. Certainly, the logistic loss is widely used in practice; are the preceding technical points at all related to the widespread empirical use of the logistic loss?

Acknowledgments and Disclosure of Funding

The authors are grateful for support from the NSF under grant IIS-1750051. MT thanks many friends for illuminating and motivating discussions: Daniel Hsu, Phil Long, Maxim Raginsky, Fanny Yang.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. arXiv:1811.04918 [cs.LG], 2018a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. arXiv:1811.03962 [cs.LG], 2018b.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. arXiv:1901.08584 [cs.LG], 2019.

Yu Bai, Song Mei, Huan Wang, and Caiming Xiong. Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification. arXiv:2102.07856 [cs.LG], 2021.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, May 1993.

Peter L. Bartlett and Mikhail Traskin. AdaBoost is consistent. *Journal of Machine Learning Research*, 8:2347–2368, 2007.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. arXiv:1906.11300 [stat.ML], 2019.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off. 2018a. arXiv:1812.11118 [stat.ML].

Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate. In *NeurIPS*, 2018b.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Alina Braun, Michael Kohler, Sophie Langer, and Harro Walk. The smoking gun: Statistical theory improves neural network estimates. 2021. arXiv:2107.09550 [math.ST].

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *NeurIPS*, 2019.

Zixiang Chen, Yuan Cao, Difan Zou, and Quanquan Gu. How much over-parameterization is sufficient to learn deep relu networks? In *ICLR*, 2021.

Lénaïc Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. arXiv:1812.07956v2 [math.OC], 2019.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *COLT*, 2020.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Kenneth R Davidson and Stanislaw J Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the geometry of Banach spaces*, volume 1, pages 317–366, 2001.

L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition*. Springer, 1996.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *ICLR*, 2019.

Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. 2017. arXiv:1703.11008 [cs.LG].

A. Farago and G. Lugosi. Strong universal consistency of neural network classifiers. *IEEE Transactions on Information Theory*, 39(4):1146–1151, 1993. doi: 10.1109/18.243433.

Gerald B. Folland. *Real analysis: modern techniques and their applications*. Wiley Interscience, 2 edition, 1999.

Spencer Frei, Yuan Cao, and Quanquan Gu. Agnostic learning of a single neuron with gradient descent. arXiv:2005.14426 [cs.LG], 2020.

K. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Netw.*, 2(3):183–192, May 1989. ISSN 0893-6080.

Surbhi Goel, Aravind Gollakota, Zhihan Jin, Sushrut Karmalkar, and Adam Klivans. Superpolynomial lower bounds for learning one-layer neural networks using gradient descent. In *ICML*, 2020a.

Surbhi Goel, Adam R. Klivans, Pasin Manurangsi, and Daniel Reichman. Tight hardness results for training depth-2 relu networks. arXiv:2011.13550 [cs.LG], 2020b.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks, 2017.

Leonid Gurvits and Pascal Koiran. Approximation and learning of convex superpositions. In Paul Vitányi, editor, *Computational Learning Theory*, pages 222–236. Springer, 1995.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531 [stat.ML], 2015.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, july 1989.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, pages 8571–8580, 2018.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. arXiv:2006.06657 [cs.LG], 2020a.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *ICLR*, 2020b.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *COLT*, pages 2109–2136, 2020a.

Ziwei Ji, Matus Telgarsky, and Ruicheng Xian. Neural tangent kernels, transportation mappings, and universal approximation. In *ICLR*, 2020b.

Olav Kallenberg. *Foundations of modern probability*. Probability and its Applications (New York). Springer-Verlag, New York, second edition, 2002.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *NeurIPS*, pages 8157–8166, 2018.

Shiyu Liang, Ruoyu Sun, and R. Srikant. Achieving small test error in mildly overparameterized neural networks. arXiv:2104.11895 [cs.LG], 2021.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*, 2020.

Preetum Nakkiran and Yamini Bansal. Distributional generalization: A new kind of generalization. 2021. arXiv:2009.08092 [cs.LG].

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. arXiv:1412.6614 [cs.LG], 2014.

Atsushi Nitanda and Taiji Suzuki. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. arXiv:1905.09870 [stat.ML], 2019.

Francesco Orabona and Dávid Pál. Parameter-free stochastic optimization of variationally coherent functions. arXiv:2102.00236 [math.OC], 2021.

Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. arXiv:1902.04674 [cs.LG], 2019.

Dominic Richards and Ilja Kuzborskij. Stability & generalisation of gradient descent for shallow neural networks without the neural tangent kernel. 2021. arXiv:2107.12723 [stat.ML].

Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. 2017. arXiv:1708.06633 [math.ST].

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Ohad Shamir. Gradient methods never overfit on separable data. arXiv:2007.00028 [cs.LG], 2020.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. 2019. arXiv:1906.03593 [cs.LG].

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. In *ICLR*, 2018.

Matus Telgarsky. Boosting with the logistic loss is consistent. In *COLT*, 2013.

Gilad Yehudai and Ohad Shamir. Learning a single neuron with gradient methods. arXiv:2001.05205 [cs.LG], 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.

Tong Zhang and Bin Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. arXiv:1811.08888 [cs.LG], 2018.

A Missing figures

Due to space limitations, Figures 1 and 2, referenced in the body, have been moved to this initial appendix section.

B Proof of Theorem 1.1 and supporting results

This appendix section proves all bounds necessary for Theorem 1.1, and also proves the consistency statement in Corollary 2.3.

B.1 Technical preliminaries

First, the key logistic loss properties.

Lemma B.1. *1. For any $a \geq b$,*

$$\frac{\phi(a)}{\phi(b)} \leq e^{a-b} \qquad \text{and} \qquad \frac{\ell(-a)}{\ell(-b)} \leq e^{a-b}.$$

In particular, for any f, g with $\sup_{\|x\|\leq 1} |f(x) - g(x)| \leq \tau$,

$$e^{-\tau}\mathcal{R}(f) \leq \mathcal{R}(g) \leq e^{\tau}\mathcal{R}(f).$$

If only $\max_k |f(x_k) - g(x_k)| \leq \tau$, then $e^{-\tau}\widehat{\mathcal{R}}(f) \leq \widehat{\mathcal{R}}(g) \leq e^{\tau}\widehat{\mathcal{R}}(f)$.

2. For any $f : \mathbb{R}^d \to \mathbb{R}$ and corresponding conditional model $\phi_f(x) := \phi(f(x))$,

$$\frac{1}{2}\left(\mathcal{R}_z(f) - \overline{\mathcal{R}}_z\right)^2 \leq 2\int (\phi_f(x) - p_y(x))^2 \,\mathrm{d}\mu_x(x) \leq \mathcal{K}_{\mathsf{bin}}(p_y, \phi_f) = \mathcal{R}(f) - \overline{\mathcal{R}}.$$

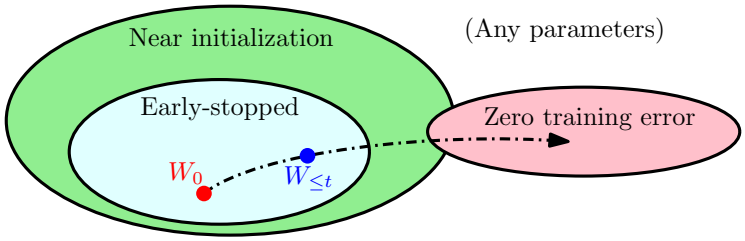
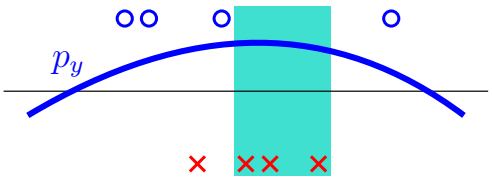


Figure 1: The setting of this paper, contrasted with standard settings. Theorem 1.1 considers iterate $W_{\leq t}$, which is somewhere in the *early-stopped* ball around the initial random choice W_0. This early-stopped ball is well inside the *near initialization* or *NTK* ball, since in noisy settings, the early-stopped ball will not reach zero training error, whereas the NTK ball will. Meanwhile, the NTK itself requires early stopping and is a subset of the space of all parameters.



(a) Conditional model p_y and some noisy data. A smoothed prediction rule would perform well.

(b) A *local interpolation rule* working very hard to fit the noisy data.

Figure 2: When data is noisy, it's best to give up on a few points. The shaded region here highlights consecutive points with the wrong label; as in Proposition 1.2, prediction rules that locally interpolate will have a large population risk in these regions.

Proof. 1. Since $a \geq b$, then $e^{b-a} \leq 1$, and

$$\frac{\phi(a)}{\phi(b)} = \frac{1+e^{-b}}{1+e^{-a}} = e^{a-b}\left(\frac{e^{b-a}+e^{-a}}{1+e^{-a}}\right) \leq e^{a-b},$$

whereby

$$\int_{-\infty}^{a} \phi(r)\,\mathrm{d}r = \int_{-\infty}^{b} \phi(r+(a-b))\,\mathrm{d}r \leq e^{a-b}\int_{-\infty}^{b} \phi(r)\,\mathrm{d}r.$$

Consequently,

$$\ell(-a) = -\int_{-a}^{\infty} \ell'(r)\,\mathrm{d}r = \int_{-a}^{\infty} \phi(-r)\,\mathrm{d}r = \int_{-\infty}^{a} \phi(r)\,\mathrm{d}r \leq e^{a-b}\int_{-\infty}^{b} \phi(r)\,\mathrm{d}r = e^{a-b}\ell(-b).$$

The first set of claims for risk follow from the fact that for any pair (x, y) and $\tau \geq 0$,

$$\ell(yr + y^2\tau) \leq \ell(yr) \leq \ell(yr - y^2\tau),$$

whereby

$$\mathcal{R}(f) = \mathbb{E}\ell(yf(x)) \leq \mathbb{E}\ell(yg(x) - \tau) \leq e^{\tau}\mathbb{E}\ell(yg(x)) = e^{\tau}\mathcal{R}(g).$$

The proof for empirical risk is similar, but only relies upon behavior on the finite sample.

2. From standard results in the literature on classification calibration [Zhang, 2004, Bartlett et al., 2006], the optimal logistic loss pointwise satisfies

$$\bar{r}_x := \inf_{r\in\mathbb{R}} p_y(x)\ell(r) + (1-p_y(x))\ell(-r) = -p_y(x)\ln p_y(x) - (1-p_y(x))\ln(1-p_y(x)).$$

Consequently, for any predictor $f : \mathbb{R}^d \to \mathbb{R}$ and corresponding probability model $\phi_f(x) := \phi(f(x))$, note that

$$\begin{aligned}\mathcal{R}(f) &= \int \left(p_y(x)\ln(1+\exp(-f(x))) + (1-p_y(x))\ln(1+\exp(f(x)))\right)\mathrm{d}\mu_x(x) \\ &= \int \left(-p_y(x)\ln\phi_f(x) - (1-p_y(x))\ln(1-\phi_f(x))\right)\mathrm{d}\mu_x(x),\end{aligned}$$

and thus

$$\mathcal{R}(f) - \overline{\mathcal{R}} = \mathcal{K}_{\text{bin}}(p_y, \phi_f).$$

By Pinsker's inequality,

$$\begin{aligned}\mathcal{K}_{\text{bin}}(p_y, \phi_f) &= \int \left(p_y(x)\ln\frac{p_y(x)}{\phi_f(x)} + (1-p_y(x))\ln\frac{1-p_y(x)}{1-\phi_f(x)}\right)\mathrm{d}\mu_x(x) \\ &\geq \frac{1}{2}\int \left(|p_y(x)-\phi_f(x)| + |(1-p_y(x)) + (1-\phi_f(x))|\right)^2\mathrm{d}\mu_x(x) \\ &= 2\int \left(p_y(x)-\phi_f(x)\right)^2\mathrm{d}\mu_x(x).\end{aligned}$$

If $\text{sgn}(\phi_f(x) - 1/2) \neq \text{sgn}(p_y(x) - 1/2)$, then $|\phi_f(x) - p_y(x)| \geq |p_y(x) - 1/2|$, and so

$$\begin{aligned}\mathcal{R}_{\text{z}}(f) - \overline{\mathcal{R}}_{\text{z}} &= \int \mathbb{1}[\text{sgn}(\phi_f(x) - 1/2) \neq \text{sgn}(p_y(x) - 1/2)] \cdot |2p_y(x) - 1|\,\mathrm{d}\mu_x(x) \\ &\leq 2\int |\phi_f(x) - p_y(x)|\,\mathrm{d}\mu_x(x) \\ &\leq 2\sqrt{\int (\phi_f(x) - p_y(x))^2\,\mathrm{d}\mu_x(x)}.\end{aligned}$$

□

The remainder of this technical subsection develops a variety of concentration inequalities used throughout, most notably the control over the sphere in Lemma B.3. First, a few standard Gaussian inequalities, included here for completeness.

Lemma B.2. *Suppose $W \in \mathbb{R}^{m\times d}$ has iid Gaussian entries $W_{j,k} \sim \mathcal{N}(0,1)$, and let $(w_j^\intercal)_{j=1}^m$ denote the rows.*

1. *For any $\tau > 0$, with probability at least $1-3\delta$,*
$$\sum_{j=1}^m \mathbb{1}\left[|w_j^\intercal x| \le \tau\|x\|\right] \le m\tau + \sqrt{8m\tau\ln(1/\delta)}.$$
2. *With probability at least $1-\delta$,*
$$\|W\|_2 < \sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)}.$$
3. *With probability at least $1-2\delta$,*
$$-\|z\|\sqrt{2\ln(1/\delta)} \le \|\sigma_{\mathrm{r}}(Wz)\| - \mathbb{E}\|\sigma_{\mathrm{r}}(Wz)\| \le \|z\|\sqrt{2\ln(1/\delta)},$$
where
$$\|z\|\left(\sqrt{\frac{m}{2}} - \frac{5}{\sqrt{8m}}\right) \le \mathbb{E}\|\sigma_{\mathrm{r}}(Wz)\| \le \|z\|\sqrt{\frac{m}{2}}.$$
4. *With probability at least $1-\delta$, $w \in \mathbb{R}^d$ with coordinates $w_i \sim \mathcal{N}(0,1)$ satisfies*
$$\|w\| \le \sqrt{d} + \sqrt{2\ln(1/\delta)}.$$

Proof. 1. For any row j, define an indicator random variable
$$P_j := \mathbb{1}[|w_j^\intercal x| \le \tau\|x\|].$$
By rotational invariance, $P_j = \mathbb{1}[|w_{j,1}| \le \tau]$, which by the form of the Gaussian density gives
$$\Pr[P_j = 1] \le \frac{2\tau}{\sqrt{2\pi}} \le \tau.$$
As such, by a multiplicative Chernoff bound [Blum et al., 2020, Theorem 12.6], with probability at least $1-3\delta$,
$$\sum_{j=1}^m P_j \le m\Pr[P_1 = 1] + \sqrt{8m\Pr[P_1=1]\ln(1/\delta)} \le m\tau + \sqrt{8m\tau\ln(1/\delta)},$$
as desired.

2. This is a standard spectral norm concentration bound for Gaussian matrices [Davidson and Szarek, 2001, Theorem II.13],

3. For the expectation, first note for a single row $w^\intercal$ by rotational invariance of the Gaussian that
$$\mathbb{E}\sigma_{\mathrm{r}}(w^\intercal x)^2 = \|x\|^2\mathbb{E}\sigma_{\mathrm{r}}(w_1)^2 = \frac{1}{2}\|x\|^2\mathbb{E}w_1^2 = \frac{\|x\|^2}{2}.$$
As such, for a full matrix W, the expected norm can be upper bounded via
$$\mathbb{E}\|\sigma_{\mathrm{r}}(Wx)\| \le \sqrt{\mathbb{E}\|\sigma_{\mathrm{r}}(Wx)\|^2} = \sqrt{\frac{1}{2}\sum_{i=1}^m \|x\|^2} = \|x\|\sqrt{m/2},$$
and by a second-order lower bound, letting $\tilde{x} = x/\|x\|$ for convenience, and dividing through by $\sqrt{m/2}$ to ease notation,
$$\begin{aligned}
\mathbb{E}\sqrt{2\|\sigma_{\mathrm{r}}(Wx)\|^2/m} &= \|x\|\mathbb{E}\sqrt{2\|\sigma_{\mathrm{r}}(W\tilde{x})\|^2/m} \\
&\ge \|x\|\mathbb{E}\left(1 + (2\|\sigma_{\mathrm{r}}(W\tilde{x})\|^2/m - 1)/2 - (2\|\sigma_{\mathrm{r}}(W\tilde{x})\|^2/m - 1)^2/2\right) \\
&= \|x\|\left(1 - \mathbb{E}(2\|\sigma_{\mathrm{r}}(W\tilde{x})\|^2/m - 1)^2/2\right) \\
&= \|x\|\left(\frac{3}{2} - \frac{m(m-1)}{2m^2} - \frac{6m}{2m^2}\right) \\
&= \|x\|\left(1 - \frac{5}{2m}\right).
\end{aligned}$$

For the concentration part, note firstly that σ_{r} is ℓ_2-Lipschitz when applied coordinate-wise, since

$$\|\sigma_{\mathrm{r}}(u) - \sigma_{\mathrm{r}}(v)\|^2 = \sum_{i=1}^m (\sigma_{\mathrm{r}}(u_i) - \sigma_{\mathrm{r}}(v_i))^2 \leq \sum_{i=1}^m (u_i - v_i)^2 = \|u - v\|^2,$$

and thus

$$\|\sigma_{\mathrm{r}}(Ax)\| - \|\sigma_{\mathrm{r}}(Bx)\| \leq \big\|\sigma_{\mathrm{r}}(Ax) - \sigma_{\mathrm{r}}(Bx)\big\| \leq \|Ax - Bx\| \leq \|A - B\| \, \|x\|,$$

and thus by standard Gaussian concentration, with probability at least $1 - \delta$,

$$\|\sigma_{\mathrm{r}}(Wx)\| - \mathbb{E}\|\sigma_{\mathrm{r}}(Wx)\| \leq \|x\| \sqrt{2\ln(1/\delta)},$$

and vice versa.

4. This is a subset of the preceding proof: $w \mapsto \|w\|$ is 1-Lipschitz, thus by standard Gaussian concentration, with probability at least $1 - \delta$,

$$\|w\| \leq \mathbb{E}\|w\| + \sqrt{2\ln(1/\delta)},$$

where $\mathbb{E}\|w\| \leq \sqrt{\mathbb{E}\|w\|^2} = \sqrt{d}$.

□

Next, finally, the control over the sphere, Lemma B.3. This lemma perhaps looks a bit underwhelming or simply abstract or overly complicated, but is a key tool in many steps of the proofs here; in particular, since it allows consideration for all $\|x\| \leq 1$, it may be applied over the distribution. This consideration over the entire sphere contrasts this lemma (and its applications) from similar inequalities in prior work [Allen-Zhu et al., 2018b, Chen et al., 2021].

Lemma B.3. *Let scalars $R_V \geq 0$, and $\epsilon \in (0, 1/(md))$, and $m \geq \ln(edm)$ be given, along with a filter set $\mathcal{S}_0 \subseteq \mathbb{R}^{m\times d}$, and define $\mathcal{S} := \mathcal{S}_0 \cap \{V \in \mathbb{R}^{m\times d} : \|V - W_0\| \leq R_V\}$. Let a function $h_V : \mathbb{R}^d \to \mathbb{R}$ be given with parameter $V \in \mathcal{S}$, and define functions*

$$\mathcal{H} := \left\{ x \mapsto h_V(x) + \big\langle \nabla f(x; W_0), V - W_0 \big\rangle : V \in \mathcal{S} \right\}.$$

Moreover, let additional scalars r_1, r_2, δ satisfy the following conditions.

1. *For every x and z with $\|x - z\| \leq \epsilon$, then $\sup_{V\in\mathcal{S}} |h_V(x) - h_V(z)| \leq r_1$.*
2. *For any fixed $\|x\| \leq 1$, with probability at least $1 - \delta$, then $\sup_{h\in\mathcal{H}} |h(x)| \leq r_2$.*

Then with probability at least $1 - (\sqrt{d}/\epsilon)^d\delta$,

$$\sup_{\|x\|\leq 1} \sup_{h\in\mathcal{H}} |h(x)| \leq r_2 + r_1 + 11R_V\rho\left(\frac{\ln(edm/\delta)}{m}\right)^{1/4}.$$

The proof of Lemma B.3 will need two technical lemmas. The first is a basic property of inner products and arccosine which also makes a later appearance in Lemma B.11.

Lemma B.4. *If $\|x - z\| \leq \epsilon$ and $x, z \neq 0$, then*

$$1 \geq \left\langle \frac{x}{\|x\|}, \frac{z}{\|z\|} \right\rangle \geq 1 - \frac{2\epsilon^2}{\|x\|^2}, \qquad \textit{and} \qquad \arccos\left(\left\langle \frac{x}{\|x\|}, \frac{z}{\|z\|} \right\rangle\right) \leq \frac{\epsilon\sqrt{8}}{\|x\|}.$$

Proof. The first inequalities follow from

$$\begin{aligned}
1 &\geq \left\langle \frac{x}{\|x\|}, \frac{z}{\|z\|} \right\rangle \\
&= 1 - \frac{1}{2}\left\| \frac{x}{\|x\|} - \frac{z}{\|z\|} \right\|^2 \\
&= 1 - \frac{1}{2\|x\|^2\|z\|^2}\left\| x\|z\| - z\|z\| + z(\|z\| - \|x\|) \right\|^2 \\
&\geq 1 - \frac{\|x - z\|^2\|z\|^2 + \|z\|^2(\|z\| - \|x\|)^2}{\|x\|^2\|z\|^2} \\
&\geq 1 - \frac{2\|x - z\|^2\|z\|^2}{\|x\|^2\|z\|^2} \\
&\geq 1 - \frac{2\epsilon^2}{\|x\|^2}.
\end{aligned}$$

To finish, since $\arccos$ is decreasing along $[0, 1]$, and since for any $a \in [0, 1]$,

$$\arccos(1 - a) = \int_{1-a}^{1} \frac{\mathrm{d}r}{\sqrt{1 - r^2}} = \int_0^a \frac{\mathrm{d}r}{\sqrt{2r - r^2}} \leq \int_0^a \frac{\mathrm{d}r}{\sqrt{r}} = 2\sqrt{a},$$

then

$$\arccos\left(\left\langle \frac{x}{\|x\|}, \frac{z}{\|z\|} \right\rangle\right) \leq \arccos\left(1 - \frac{2\epsilon^2}{\|x\|^2}\right) \leq 2\sqrt{\frac{2\epsilon^2}{\|x\|^2}} = \frac{\epsilon\sqrt{8}}{\|x\|}.$$

□

The main heavy lifting in Lemma B.3 is encapsulated in the following concentration inequality. In words, it controls the behavior of the initial features within a tiny localized region of the sphere; the proof of Lemma B.3 combines this local control with a discrete cover of the sphere, together giving control over the entire sphere.

Lemma B.5. *Let any fixed $\|z\| \leq 1$ be given (independent of W_0), along with a scalar $\epsilon > 0$ with $\epsilon \leq 1/(dm)$, where $m \geq \ln(edm)$. Then, with probability at least $1 - \delta$,*

$$\sup_{\substack{\|x-z\| \leq \epsilon \\ \|x\| \leq 1}} \|\nabla f(x; W_0) - \nabla f(z; W_0)\|^2 \leq 113\rho^2\sqrt{\frac{\ln(edm/\delta)}{m}}.$$

Proof. Throughout the proof, simplify notation via $W := W_0$, and let $(w_j^\intercal)_{j=1}^m$ denote the rows of W, and furthermore write

$$g(x, z; w) := \frac{\rho^2}{m}\left\| x\mathbb{1}[w_j^\intercal x \geq 0] - z\mathbb{1}[w_j^\intercal z \geq 0] \right\|^2.$$

Lastly, for any $x \in \mathbb{R}^d$ under consideration, then $\|x\| \leq 1$, so this condition will often be implicit. Note that

$$\begin{aligned}
\sup_{\|x-z\| \leq \epsilon} \|\nabla f(x; W) - \nabla f(z; W)\|^2 &= \frac{\rho^2}{m} \sup_{\|x-z\| \leq \epsilon} \sum_{j=1}^m \|x\mathbb{1}[w_j^\intercal x \geq 0] - z\mathbb{1}[w_j^\intercal z \geq 0]\|^2 \\
&= \sup_{\|x-z\| \leq \epsilon} \sum_{j=1}^m g(x, z; w_j).
\end{aligned}$$

Next note that this quantity, treated as a function of the m rows of W, satisfies bounded differences with constant ρ^2/m: letting W' be a copy of W which differs only in a single row w_i', and noting

$g \geq 0$,

$$\left| \sup_{\|x-z\|\leq\epsilon} \sum_{j=1}^{m} g(x,z;w_j) - \sup_{\|x-z\|\leq\epsilon} \sum_{j=1}^{m} g(x,z;w'_j) \right|$$

$$= \left| \sup_{\|x-z\|\leq\epsilon} \sum_{j=1}^{m} g(x,z;w_j) - \sup_{\|x-z\|\leq\epsilon} \left(g(x,z;w_i) - g(x,z;w_i) + \sum_{j=1}^{m} g(x,z;w'_j) \right) \right|$$

$$\leq \sup_{\|x-z\|\leq\epsilon} \left| g(x,z;w'_i) - g(x,z;w_i) \right| \leq \frac{\rho^2}{m}.$$

As such, by McDiarmid's inequality, with probability at least $1-\delta$,

$$\sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2 \leq \sqrt{\rho^4 \ln(1/\delta)/(2m)} + \mathbb{E}_W \sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2. \tag{2}$$

It remains to analyze this expectation. First consider the case that $\|z\| \leq 3\sqrt{\epsilon}$; then, for any W,

$$\begin{aligned}
\sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2 &= \frac{\rho^2}{m} \sup_{\|x-z\|\leq\epsilon} \sum_{j=1}^{m} \|x\mathbb{1}[w_j^\intercal x \geq 0] - z\mathbb{1}[w_j^\intercal z \geq 0]\|^2 \\
&\leq \frac{2\rho^2}{m} \sup_{\|x-z\|\leq\epsilon} \sum_{j=1}^{m} \left(\|x\|^2 + \|z\|^2 \right) \\
&\leq \frac{2\rho^2}{m} \sum_{j=1}^{m} (16\epsilon + 9\epsilon) \leq 50\epsilon\rho^2.
\end{aligned} \tag{3}$$

For the rest of the proof, suppose $\|z\| > 3\sqrt{\epsilon}$, which also implies $\|x\| > 2\sqrt{\epsilon}$ for every x satisfying $\|x-z\| \leq \epsilon$.

Since z is fixed, and in particular does not depend on W, we may use the rotational invariance of W to leverage the condition $\|x-z\| \leq \epsilon$. Specifically, define a matrix $M \in \mathbb{R}^{d\times d}$ whose first column is $z/\|z\|$, and the remaining columns are orthonormal (we can not use x in the definition of M, since x varies within the expectation). Defining (for any x) the two projections $x_z := zx^\intercal z/\|z\|^2$ and $x^\perp := x - x_z$ (whereby $z^\intercal x^\perp = 0$), we may rotate the rows of W by M, giving

$$\begin{aligned}
\mathbb{1}\left[(Mw_j)^\intercal z \geq 0\right] &= \mathbb{1}\left[w_{j,1}\|z\| \geq 0\right] \\
&= \mathbb{1}\left[w_{j,1} \geq 0\right], \\
\mathbb{1}\left[(Mw_j)^\intercal x \geq 0\right] &= \mathbb{1}\left[w_j^\intercal M^\intercal (x_z + x^\perp) \geq 0\right] \\
&= \mathbb{1}\left[w_{j,1} z^\intercal x/\|z\| \geq -w_j^\intercal M^\intercal x^\perp\right] \\
&= \mathbb{1}\left[w_{j,1} \geq -\frac{w_j^\intercal M^\intercal x^\perp}{z^\intercal x/\|z\|}\right],
\end{aligned}$$

where the last division does not change the sign due to $\|z-x\| \leq \epsilon$ and $\|z\| > 3\sqrt{\epsilon}$, for instance as verified by upcoming invocations of Lemma B.4. Now let E_j denote the event that for this w_j, there exists $\|x-z\| \leq \epsilon$ such that these two indicators are not equal. Letting $\tau > 0$ denote a free parameter to be optimized later, this event is implied by the union of two simpler events: let $w_{j,2:} \in \mathbb{R}^{d-1}$ denote all but the first coordinate of w_j, and define

$$E_{j,1} := \left[|w_{j,1}| \leq \tau\right], \qquad E_{j,2} := \left[\sup_{\|x-z\|\leq\epsilon} \frac{\|w_{j,2:}\| \cdot \|x^\perp\| \cdot \|z\|}{z^\intercal x} > \tau\right];$$

by construction (and Cauchy-Schwarz), if the negation of both events holds, then the indicators are the same. To upper bound the probability of the first event, by the form of the Gaussian density,

$$\Pr[E_{j,1}] \leq \tau\sqrt{\frac{2}{\pi}} < \tau.$$

To control the various terms in $E_{j,2}$, firstly by Lemma B.2, with probability at least $1-\epsilon$, then

$$\|w_{j,2:}\| \leq \sqrt{d-1} + \sqrt{2\ln(1/\epsilon)} \leq \sqrt{2d - 2 + 4\ln(1/\epsilon)};$$

this will be the only step of the derivation controlling $\Pr[E_{j,2}]$, and note that it depends only on w_j and z and not on any specific x. Next, by Lemma B.4, for any $\|x-z\| \leq \epsilon$, since $\|x\| \geq 2\epsilon$ (whereby $2\epsilon^2/\|x\|^2 < 1$),

$$\begin{aligned}
\|x^\perp\|^2 = \|x\|^2 - \frac{(z^\mathsf{T} x)^2}{\|z\|^2} &= \|x\|^2 \left(1 - \left[\frac{z^\mathsf{T} x}{\|x\|\|z\|}\right]^2\right) \\
&\leq \|x\|^2 \left(1 - \left[1 - \frac{2\epsilon^2}{\|x\|^2}\right]^2\right) = 4\epsilon^2 - \frac{4\epsilon^4}{\|x\|^2} \leq 4\epsilon^2.
\end{aligned}$$

Similarly by Lemma B.4, using $\epsilon \leq 1$,

$$\frac{z^\mathsf{T} x}{\|z\|} \geq \|x\| - \frac{2\epsilon^2}{\|x\|} > 2\sqrt{\epsilon} - \epsilon^{1.5} \geq \sqrt{\epsilon}.$$

Combining all these pieces, with probability at least $1-\epsilon$,

$$\frac{\|w_{j,2:}\| \cdot \|x^\perp\| \cdot \|z\|}{z^\mathsf{T} x} \leq \sqrt{2d - 2 + 4\ln(1/\epsilon)} \left(\frac{2\epsilon}{\sqrt{\epsilon}}\right) \leq 4\sqrt{d\epsilon \ln(e/\epsilon)}.$$

This right hand side does not depend on the specific choice of x, and holds for any $\|x-z\| \leq \epsilon$. As such, set $\tau := 4\sqrt{d\epsilon\ln(e/\epsilon)}$, whereby

$$\Pr[E_j] \leq \Pr[E_{j,1}] + \Pr[E_{j,2}] \leq \tau + \epsilon.$$

Moreover, by a multiplicative Chernoff bound [Blum et al., 2020, Theorem 12.6], with probability at least $1-3\epsilon$, the events $(E_j)_{j=1}^m$ hold for at most $m_\tau := m(\tau+\epsilon) + \sqrt{8m(\tau+\epsilon)\ln(1/\epsilon)}$ rows. Now let E_τ denote the event that $(E_j)_{j=1}^m$ holds for at most m_τ rows. Then

$$\begin{aligned}
&\mathbb{E}_W \sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2. \\
&= \mathbb{E}_W \left[\sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2 \mid E_\tau\right] \Pr[E_\tau] \\
&\quad + \mathbb{E}_W \left[\sup_{\|x-z\|\leq\epsilon} \|\nabla f(x;W) - \nabla f(z;W)\|^2 \mid E_\tau^c\right] \Pr[E_\tau^c] \\
&\leq 2\rho^2 \sup_{\|x-z\|\leq\epsilon} \left(\frac{m}{m}\|x-z\|^2 + \frac{m_\tau}{m}(\|x\|^2 + \|z\|^2)\right) + \sup_{\|x-z\|\leq\epsilon} \left(3\epsilon(\|x\|^2 + \|z\|^2)\right) \\
&\leq 2\rho^2 \left(\epsilon^2 + \frac{2m_\tau}{m} + 6\epsilon\right). \qquad (4)
\end{aligned}$$

The proof will now be completed by returning to the McDiarmid application resulting in eq. (2), and combining all preceding bounds. Starting with a simplification via the assumption $\epsilon \leq 1/(dm)$ and $m \geq \ln(edm)$, note

$$\begin{aligned}
\tau = 4\sqrt{d\epsilon\ln(e/\epsilon)} &\leq 4\sqrt{\frac{\ln(edm)}{m}}, \\
\frac{m_\tau}{m} &= \tau + \epsilon + \sqrt{8(\tau+\epsilon)\ln(1/\epsilon)/m} \\
&\leq 5\sqrt{\frac{\ln(edm)}{m}} + \sqrt{\frac{40\sqrt{\ln(edm)}\ln(edm)}{m^{3/2}}} \leq 12\sqrt{\frac{\ln(edm)}{m}}.
\end{aligned}$$

Combining the preceding simplifications with eqs. (3) and (4), continuing from the McDiarmid application in eq. (2), with probability at least $1-\delta$,

$$\begin{aligned}\sup_{\substack{\|x-z\|\le\epsilon\\ \|x\|\le 1}} \|\nabla f(x;W_0)-\nabla f(z;W_0)\|^2 &\le \rho^2\left(\sqrt{\frac{\ln(1/\delta)}{2m}}+50\epsilon+2\left(\epsilon^2+\frac{2m_\tau}{m}+6\epsilon\right)\right)\\ &\le \rho^2\left(\sqrt{\frac{\ln(1/\delta)}{2m}}+\frac{50}{md}+\frac{2}{m^2d^2}+48\sqrt{\frac{\ln(edm)}{m}}+\frac{12}{md}\right)\\ &\le 113\rho^2\sqrt{\frac{\ln(edm/\delta)}{m}}.\end{aligned}$$

□

Finally, the proof of Lemma B.3 via the preceding technical lemmas.

Proof of Lemma B.3. Let $\mathcal{C}$ denote a cover of each coordinate of $\|x\|\le 1$ at scale $\epsilon/\sqrt{d}$, meaning $|\mathcal{C}|\le(\sqrt{d}/\epsilon)^d$ (the grid elements can be $2\epsilon/\sqrt{d}$ apart), and for any $\|x\|\le 1$, there exists $z\in\mathcal{C}$ with

$$\|z-x\| = \sqrt{\sum_{i=1}^d (z_i-x_i)^2}\le\epsilon.$$

This cover $\mathcal{C}$ will be used throughout the proof; it is crucial that its construction makes no reference to W_0, and in particular that the cover elements are independent of W_0.

Union bound together and discard $|\mathcal{C}|\delta$ failure probability so that for every $z\in\mathcal{C}$, then $\sup_{h\in\mathcal{H}}|h(z)|\le r_2$. Additionally union bound together and discard $|\mathcal{C}|\delta$ failure probability corresponding to instantiating Lemma B.5 for each $z\in\mathcal{C}$, whereby

$$\max_{z\in\mathcal{C}}\sup_{\substack{\|x-z\|\le\epsilon\\ \|x\|\le 1}}\|\nabla f(x;W_0)-\nabla f(z;W_0)\|^2\le 113\rho^2\sqrt{\frac{\ln(edm/\delta)}{m}}.$$

Now let an arbitrary $\|x\|\le 1$ be given, and let $z\in\mathcal{C}$ be a nearest cover element, whereby $\|z-x\|\le\epsilon$. Then

$$\begin{aligned}\sup_{h\in\mathcal{H}}|h(x)| &\le \sup_{h\in\mathcal{H}}\Big(|h(z)|+|h(z)-h(x)|\Big)\\ &\le r_2+\sup_{V\in\mathcal{S}}|h_V(z)-h_V(x)|+\sup_{V\in\mathcal{S}}|\left\langle\nabla f(x;W_0)-\nabla f(z;W_0),V-W_0\right\rangle|\\ &\le r_2+r_1+\sup_{V\in\mathcal{S}}\big\|\nabla f(x;W_0)-\nabla f(z;W_0)\big\|\cdot\|V-W_0\|\\ &\le r_2+r_1+11R_V\rho\left(\frac{\ln(edm/\delta)}{m}\right)^{1/4}.\end{aligned}$$

□

As a first application of Lemma B.3, the range of the mappings can be bounded for all $\|x\|\le 1$, which is used later in the generalization analysis.

Lemma B.6. *Let $R_V>0$ be given.*

1. *For any $x\in\mathbb{R}^d$, with probability at least $1-3\delta$, every $V\in\mathbb{R}^{m\times d}$ satisfies*
$$\Big|\big\langle\nabla f(x;W_0),V\big\rangle\Big|\le\rho\|x\|\left(\|V-W_0\|_{\mathrm{F}}+2\ln(1/\delta)\right).$$
2. *Suppose $R_V\ge 1$ and $m\ge\ln(emd)$. With probability at least $1-(1+3(md^{3/2})^d)\delta$,*
$$\sup_{\|V-W_0\|\le R_V}\sup_{\|x\|\le 1}\Big|\big\langle\nabla f(x;W_0),V\big\rangle\Big|\le 18R_V\rho\ln(emd/\delta).$$

Proof. For convenience throughout the proof, write $W := W_0$.

1. Splitting terms via $V = V - W + W$,
$$\left|\left\langle \nabla f(x; W), V \right\rangle\right| \leq \left|\left\langle \nabla f(x; W), W \right\rangle\right| + \left|\left\langle \nabla f(x; W), V - W \right\rangle\right|.$$
For the first term, since W is independent of a and can be treated as fixed, by Hoeffding's inequality, with probability at least $1 - 2\delta$ over the draw of a,
$$\left|\left\langle \nabla f(x; W), W \right\rangle\right| = \left|f(x; W)\right| \leq \frac{\rho}{\sqrt{m}} \|\sigma_{\mathrm{r}}(Wx)\| \sqrt{\ln(1/\delta)/2}.$$
By Lemma B.2, with additional failure probability δ,
$$\|\sigma_{\mathrm{r}}(Wx)\| \leq \mathbb{E}\|\sigma_{\mathrm{r}}(Wx)\| + \|x\|\sqrt{2\ln(1/\delta)} \leq \|x\| \left(\sqrt{m/2} + \sqrt{2\ln(1/\delta)}\right).$$
Together,
$$\left|\left\langle \nabla f(x; W), W \right\rangle\right| \leq \rho\|x\| \left(1 + \sqrt{2\ln(1/\delta)/m}\right) \sqrt{\ln(1/\delta)/2}.$$
For the second term, due to the scale of the first term, it suffices to worst-case everything: by Cauchy-Schwarz,
$$\left|\left\langle \nabla f(x; W), V - W \right\rangle\right| \leq \|\nabla f(x; W)\|_{\mathrm{F}} \cdot \|V - W\|_{\mathrm{F}} \leq \rho\|x\| \cdot \|V - W\|_{\mathrm{F}}.$$
Combining everything, with probability at least $1 - 3\delta$,
$$\begin{aligned}\left|\left\langle \nabla f(x; W), V \right\rangle\right| &\leq \rho\|x\| \left(\|V - W\|_{\mathrm{F}} + \sqrt{\ln(1/\delta)/2} + \ln(1/\delta)/\sqrt{m}\right) \\ &\leq \rho\|x\| \left(\|V - W\|_{\mathrm{F}} + 2\ln(1/\delta)\right)\end{aligned}$$

2. This item proceeds by combining the previous item with the covering argument from Lemma B.3. Concretely, define the function
$$h_V(x) := f^{(0)}(x);$$
that is, h_V has no dependence on $V \in \mathbb{R}^{m \times d}$, but note that
$$\left\langle \nabla f(x; W), V \right\rangle = \left\langle \nabla f(x; W), V - W \right\rangle + \left\langle \nabla f(x; W), W \right\rangle = \left\langle \nabla f(x; W), V - W \right\rangle + h_V(x),$$
which is precisely the expression controlled by Lemma B.3. Let $\mathcal{H}$ denote the class of functions defined there.

By the preceding item, for any fixed $\|x\| \leq 1$, with probability at least $1 - \delta$,
$$\sup_{h \in \mathcal{H}} |h(x)| \leq \rho\left(R_V + 2\ln(1/\delta)\right) =: r_2.$$
Moreover, by Lemma B.2, with probability at least $1-\delta$, then $\|W\|_2 \leq \sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)}$, thus for any $\|x - z\| \leq \epsilon$, with ϵ to be determined later,
$$\begin{aligned}|f(x; W) - f(z; W)| &\leq \frac{\rho}{\sqrt{m}} \|a\| \cdot \|W(x - z)\| \leq \rho\|W\|_2 \|x - z\| \\ &\leq \rho(\sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)})\epsilon =: r_1.\end{aligned}$$
As such, by Lemma B.3, choosing $\epsilon := 1/(md)$ and $\mathcal{S}_0 = \mathbb{R}^{m \times d}$, with probability at least $1 - 3(md^{3/2})^d\delta$,
$$\begin{aligned}\sup_{\|V - W\| \leq R_V} \sup_{\|x\| \leq 1} h_V(x) &\leq r_2 + r_1 + 11R_V\rho\left(\frac{\ln(edm/\delta)}{m}\right)^{1/4}. \\ &\leq \rho\left(R_V + 2\ln(1/\delta)\right) \\ &\quad + \rho(\sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)})\epsilon \\ &\quad + 11R_V\rho\left(\frac{\ln(edm/\delta)}{m}\right)^{1/4}. \\ &\leq 18R_V\rho\ln(emd/\delta).\end{aligned}$$

□

Next, the linear approximation bounds; the last two items use Lemma B.3 to control all points on the sphere. As mentioned before, this is in contrast to prior presentations of linear approximation inequalities, which only establish the bounds on the finite training sample [Chen et al., 2021, Allen-Zhu et al., 2018b]. Note that the bounds over the sphere have a more restrictive statement; the present proof does not handle the more general form presented for a finite sample.

Lemma B.7 (See also Lemma 2.1)**.** *Let scalars $\delta > 0$ and $R_V \geq 1$ and $R_B \geq 0$ be given.*

1. *For any fixed $x \in \mathbb{R}^d$, with probability at least $1 - 3\delta$, for any $V \in \mathbb{R}^{m\times d}$ and $B \in \mathbb{R}^{m\times d}$ with $\|V - W_0\| \leq R_V$ and $\|B - W_0\| \leq R_B$,*

$$\left|\left\langle \nabla f(x; V) - \nabla f(x; W_0), B\right\rangle\right| \leq \frac{3\rho\|x\|\left(R_B + 2R_V\right) R_V^{1/3} \ln(e/\delta)^{1/4}}{m^{1/6}} =: \tau_1.$$

2. *Let τ_1 be as in the previous part. With probability at least $1 - 3n\delta$,*

$$\sup_{\|W_i - W_0\|\leq R_V} \sup_{\|W_j - W_0\|\leq R_V} \sup_{\|B - W_0\|\leq R_B} \frac{\widehat{\mathcal{R}}^{(i)}(B)}{\widehat{\mathcal{R}}^{(j)}(B)} \leq e^{2\tau_1}.$$

3. *Suppose $m \geq \ln(edm)$. With probability at least $1 - (1 + 3(d^2m)^d)\delta$,*

$$\sup_{\|V - W_0\|\leq R_V} \sup_{\|x\|\leq 1} \left|\left\langle \nabla f(x; V) - \nabla f(x; W_0), V\right\rangle\right| \leq \frac{25\rho R_V^{4/3}\sqrt{\ln(edm/\delta)}}{m^{1/6}} =: \tau_3.$$

4. *Let τ_3 be as in the previous part and again suppose $m \geq \ln(edm)$. With probability at least $1 - (1 + 3(d^2m)^d)\delta$,*

$$\sup_{\|W_i - W_0\|\leq R_V} \frac{\mathcal{R}(W_i)}{\mathcal{R}^{(0)}(W_i)} \leq e^{\tau_3}.$$

Proof of Lemmas 2.1 and B.7. The first item implies the second via Lemma B.1, and moreover implies the third item via Lemma B.3. Similarly, the third item implies the fourth via Lemma B.1. Throughout the proof, write $W := W_0$ with rows $(w_j^{\mathsf{T}})_{j=1}^m$ for convenience.

1. Fix $x \in \mathbb{R}^d$. Fix a parameter $r > 0$, which will be optimized at the end of the proof. Let V and B be given with $\|V - W\| \leq R_V$ and $\|B - W\| \leq R_B$.

   Define the sets

$$\begin{aligned} S_1 &:= \left\{j \in [m] : |w_j^{\mathsf{T}} x| \leq r\|x\|\right\}, \\ S_2 &:= \left\{j \in [m] : \|v_j - w_j\| \geq r\right\} \\ S &:= S_1 \cup S_2. \end{aligned}$$

   By Lemma B.2, with probability at least $1 - 3\delta$,

$$|S_1| \leq rm + \sqrt{8rm\ln(1/\delta)}.$$

   On the other hand,

$$R_V^2 \geq \|V - W\|^2 \geq \sum_{j\in S_2} \|v_j - w_j\|^2 \geq |S_2| r^2,$$

   meaning $|S_2| \leq R_V^2/r^2$. For any $j \notin S$, if $w_j^{\mathsf{T}} x > 0$, then

$$v_j^{\mathsf{T}} x \geq w_j^{\mathsf{T}} x - \|v_j - w_j\| \cdot \|x\| > \|x\|\,(r - r) = 0,$$

   meaning $\mathbb{1}[w_j^{\mathsf{T}} x \geq 0] = \mathbb{1}[v_j^{\mathsf{T}} x \geq 0]$; the case that $j \notin S$ and $w_j^{\mathsf{T}} x < 0$ is analogous. Together,

$$|S| \leq rm + \sqrt{8rm\ln(1/\delta)} + \frac{R_V^2}{r^2} \quad \text{and} \quad j \notin S \Longrightarrow \mathbb{1}[w_j^{\mathsf{T}} x \geq 0] = \mathbb{1}[v_j^{\mathsf{T}} x \geq 0].$$

Continuing,

$$\begin{aligned}
&\frac{\sqrt{m}}{\rho}\left|\left\langle \nabla f(x;V) - \nabla f(x;W), B\right\rangle\right| \\
&\le \frac{\sqrt{m}}{\rho}\left|\left\langle \nabla f(x;V) - \nabla f(x;W), V\right\rangle\right| + \frac{\sqrt{m}}{\rho}\left|\left\langle \nabla f(x;V) - \nabla f(x;W), V-B\right\rangle\right| \\
&= \left|a^\tau \left(\mathrm{diag}(\mathbb{1}[V^\tau x \ge 0]) - \mathrm{diag}(\mathbb{1}[W^\tau x \ge 0])\right) Vx\right| \\
&\quad + \left|a^\tau \left(\mathrm{diag}(\mathbb{1}[V^\tau x \ge 0]) - \mathrm{diag}(\mathbb{1}[W^\tau x \ge 0])\right) (V-B)x\right|.
\end{aligned}$$

Handling these two terms separately, the second term is easier: by Cauchy-Schwarz,

$$\begin{aligned}
\left|a^\tau \left(\mathrm{diag}(\mathbb{1}[V^\tau x \ge 0]) - \mathrm{diag}(\mathbb{1}[W^\tau x \ge 0])\right) (V-B)x\right| &\le \sqrt{|S|}\left\|(V - W - (B-W))x\right\| \\
&\le \sqrt{|S|}\left(R_V + R_B\right)\|x\|.
\end{aligned}$$

For the first term,

$$\left|a^\tau \left(\mathrm{diag}(\mathbb{1}[V^\tau x \ge 0]) - \mathrm{diag}(\mathbb{1}[W^\tau x \ge 0])\right) Vx\right| \le \sum_{j=1}^m \mathbb{1}[\mathrm{sgn}(v_j^\tau x) \ne \mathrm{sgn}(w_j^\tau x)] \cdot |v_j^\tau x|.$$

If $v_j^\tau x$ and $w_j^\tau x$ have different signs, then $|v_j^\tau x| \le |v_j^\tau x - w_j^\tau x| \le \|v_j - w_j\| \cdot \|x\|$; plugging this in, by Cauchy-Schwarz,

$$\begin{aligned}
\sum_{j=1}^m \mathbb{1}[\mathrm{sgn}(v_j^\tau x) \ne \mathrm{sgn}(w_j^\tau x)] \cdot |v_j^\tau x| &\le \sum_{j=1}^m \mathbb{1}[\mathrm{sgn}(v_j^\tau x) \ne \mathrm{sgn}(w_j^\tau x)] \cdot \|v_j - w_j\| \cdot \|x\| \\
&\le \sum_{j \in S} \|v_j - w_j\| \cdot \|x\| \\
&\le \sqrt{|S|}\|V - W\|_{\mathrm{F}}\|x\| \\
&\le R_V\sqrt{|S|}\|x\|.
\end{aligned}$$

Combining these derivations,

$$\begin{aligned}
\left|\left\langle \nabla f(x;V) - \nabla f(x;W), B\right\rangle\right| &\le \frac{\rho}{\sqrt{m}}\left(\sqrt{|S|}\left(R_V + R_B\right)\|x\| + R_V\sqrt{|S|}\|x\|\right) \\
&\le \frac{\rho\sqrt{|S|}\|x\|\left(2R_V + R_B\right)}{\sqrt{m}}.
\end{aligned}$$

Rearranging, and expanding the definition of $|S|$ with the choice $r := R_V^{2/3} m^{-1/3}$, and using $R_V \ge 1$,

$$\begin{aligned}
\left|\left\langle \nabla f(x;V) - \nabla f(x;W), B\right\rangle\right| &\le \frac{\rho\|x\|\left(R_B + 2R_V\right)}{\sqrt{m}}\sqrt{rm + \sqrt{8rm\ln(1/\delta)} + \frac{R_V^2}{r^2}} \\
&\le \frac{\rho\|x\|\left(R_B + 2R_V\right)R_V^{1/3}m^{1/3}\ln(e/\delta)^{1/4}}{\sqrt{m}}\sqrt{1 + \sqrt{8} + 1} \\
&\le \frac{3\rho\|x\|\left(R_B + 2R_V\right)R_V^{1/3}\ln(e/\delta)^{1/4}}{m^{1/6}}.
\end{aligned}$$

2. Union bounding the previous part over all $(x_k)_{k=1}^n$, with probability at least $1-\delta$, for any iterations (i,j) and for any matrices (W_i, W_j, B) satisfying $\max\{\|W_i - W_0\|, \|W_j - W_0\|, \|B - W_0\|\} \le R_V$

$$\max_k \left|\left\langle \nabla f(x_k;W_i) - \nabla f(x_k;W), B\right\rangle\right| \le \tau_1.$$

In particular, by Lemma B.1,

$$e^{-\tau_1} \le \frac{\widehat{\mathcal{R}}^{(i)}(B)}{\widehat{\mathcal{R}}^{(0)}(B)} \le e^{\tau_1}.$$

Applying this twice gives

$$e^{-2\tau_1} \leq \frac{\widehat{\mathcal{R}}^{(i)}(B)}{\widehat{\mathcal{R}}^{(0)}(B)} \left(\frac{\widehat{\mathcal{R}}^{(0)}(B)}{\widehat{\mathcal{R}}^{(j)}(B)} \right) = \frac{\widehat{\mathcal{R}}^{(i)}(B)}{\widehat{\mathcal{R}}^{(j)}(B)} \leq e^{2\tau_1}.$$

3. This part follows from the first via Lemma B.3. As such, for every $\|V - W\| \leq R_V$, define

$$h_V(x) := f(x; W) - f(x; V);$$

by this choice,

$$\begin{aligned}
\left\langle \nabla f(x; V) - \nabla f(x; W), V \right\rangle &= \left\langle \nabla f(x; V), V \right\rangle - \left\langle \nabla f(x; W), W \right\rangle - \left\langle \nabla f(x; W), V - W \right\rangle \\
&= f(x; V) - f(x; W) - \left\langle \nabla f(x; W), V - W \right\rangle \\
&= -h_V(x) - \left\langle \nabla f(x; W), V - W \right\rangle,
\end{aligned}$$

which matches the (negation of) functions considered in the function class $\mathcal{H}$ in Lemma B.3.

By the previous part, with $R_B := 0$, for any fixed $\|x\| \leq 1$, with probability at least $1 - 3\delta$,

$$\sup_{h \in \mathcal{H}} |h(x)| \leq \frac{6\rho R_V^{4/3} \ln(e/\delta)^{1/4}}{m^{1/6}} =: r_2.$$

Next, with probability at least $1 - \delta$, Lemma B.2 gives

$$\|W\|_2 \leq \sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)},$$

and thus for any $\|x - z\| \leq \epsilon$, since the ReLU is 1-Lipschitz even when applied to vectors,

$$\begin{aligned}
|h_V(x) - h_V(z)| &\leq |f(x; V) - f(z; V)| + |f(x; W) - f(z; W)| \\
&\leq \rho \|(V - W + W)(x - z)\| + \rho \|W(x - z)\| \\
&\leq 2\rho\epsilon(R_V/2 + \sqrt{m} + \sqrt{d} + \sqrt{2\ln(1/\delta)}) =: r_1.
\end{aligned}$$

Together, by Lemma B.3, choosing $\epsilon := 1/(dm)$ and $\mathcal{S}_0 := \mathbb{R}^{m \times d}$, with probability at least $1 - (1 + 3(md^{3/2})^d)\delta$,

$$\sup_{\|x\| \leq 1} \sup_{h \in \mathcal{H}} |h(x)| \leq r_2 + r_1 + 11 R_V \rho \left(\frac{\ln(edm/\delta)}{m} \right)^{1/4} \leq \frac{25\rho R_V^{4/3} \sqrt{\ln(edm/\delta)}}{m^{1/6}}.$$

4. By the previous item, with probability at least $1 - (1 + 3(md^{3/2})^d)\delta$,

$$\sup_{\|W_i - W_0\| \leq R_V} \sup_{\|x\| \leq 1} \left| f^{(0)}(x; W_i) - f(x; W_i) \right| \leq \tau_3.$$

Consequently, by Lemma B.1, for any W_i with $\|W_i - W_0\| \leq R_V$,

$$\mathcal{R}(W_i) = \mathbb{E}_{x,y} \ell(y f(x; W_i)) \leq e^{\tau_3} \mathbb{E}_{x,y} \ell(y f^{(0)}(x; W_i)) = e^{\tau_3} \mathcal{R}^{(0)}(W_i).$$

□

B.2 Generalization proofs

As mentioned before, the usual hard part of such a proof is the Rademacher complexity estimate, but here it is easy: linear predictors, as this bound is applied after linear approximation. The difficult step is to control the range, which was presented before in Lemma B.6, which invokes the sphere control technique in Lemma B.3.

Lemma B.8. *Let $R_V \geq 1$ and $m \geq \ln(edm)$ be given. With probability at least $1 - 6\delta$,*

$$\sup_{\|V - W_0\| \leq R_V} \mathcal{R}^{(0)}(V) - \widehat{\mathcal{R}}^{(0)}(V) \leq \frac{80\rho R_V \left(d \ln(em^2 d^3/\delta) \right)^{3/2}}{\sqrt{n}}.$$

Similarly, the negation of this bound holds with probability at least $1 - 6\delta$.

Proof. This proof will use a constant δ_0, chosen at the end. First note that the Rademacher complexity is as for linear predictors:

$$\begin{aligned}
n\text{Rad}\left(\left\{x \mapsto \left\langle \nabla f(x; W_0), V\right\rangle : \|V - W_0\| \le R_V\right\}\right) &= \mathbb{E}_\epsilon \sup_{V\in\mathcal{V}} \sum_{k=1}^n \epsilon_k \left\langle \nabla f(x_k; W_0), V\right\rangle \\
&= \mathbb{E}_\epsilon \sup_{V\in\mathcal{V}} \sum_{k=1}^n \epsilon_k \left\langle \nabla f(x_k; W_0), V - W_0 + W_0\right\rangle \\
&= \mathbb{E}_\epsilon \sup_{V\in\mathcal{V}} \sum_{k=1}^n \epsilon_k \left\langle \nabla f(x_k; W_0), V - W_0\right\rangle \\
&\le \|V - W_0\|_{\text{F}} \sqrt{\sum_{k=1}^n \|\nabla f(x_k; W_0)\|^2} \\
&\le \rho R_V \sqrt{n}.
\end{aligned}$$

Next, by Lemma B.6, with probability at least $1 - (1 + 3(md^2)^d)\delta_0$, the mappings $(x, y) \mapsto \ell(yf^{(0)}(x; V))$ are nonnegative, centered at $\ell(0)$, and vary by at most $18\rho R_V \ln(emd/\delta_0)$, thus take their amplitude to be $36\rho R_V \ln(emd/\delta_0)$ for simplicity. As such, since ℓ is 1-Lipschitz, by a standard Rademacher bound [Shalev-Shwartz and Ben-David, 2014], with additional failure probability at most $2\delta_0$,

$$\begin{aligned}
\sup_{\|V-W_0\|\le R_V} \mathcal{R}^{(0)}(V) - \widehat{\mathcal{R}}^{(0)}(V) &\le \frac{2\rho R_V}{\sqrt{n}} + \frac{108\rho R_V \ln(emd/\delta_0)\sqrt{\ln(1/\delta_0)}}{\sqrt{2n}} \\
&\le \frac{80\rho R_V \ln(emd/\delta_0)^{3/2}}{\sqrt{n}},
\end{aligned}$$

and the bound is complete by noting the total failure probability was at most $(3 + 3(md^2)^d)\delta_0 \le 6(md^2)^d\delta_0$, and setting $\delta_0 := \delta/(md^2)^d$ and simplifying.

For the reverse inequality, it follows by negating every element in the loss class and repeating the proof. □

B.3 Optimization proofs

First, a smoothness inequality which fixes the feature mapping across a pair of iterates. This lemma doesn't seem to have appeared before, but is not necessarily an improvement, other than allowing slightly larger step sizes.

Lemma B.9. *For any step size $\eta \ge 0$,*

$$\eta(1 - \eta\rho^2/8)\|\nabla\widehat{\mathcal{R}}(W_i)\|^2 \le \widehat{\mathcal{R}}^{(i)}(W_i) - \widehat{\mathcal{R}}^{(i)}(W_{i+1}).$$

If $\eta \le 8/\rho^2$, then $\widehat{\mathcal{R}}^{(i)}(W_{i+1}) \le \widehat{\mathcal{R}}^{(i)}(W_i)$, and any choice $\eta \le 4/\rho^2$ grants

$$\frac{\eta}{2}\|\nabla\widehat{\mathcal{R}}(W_i)\|^2 \le \widehat{\mathcal{R}}^{(i)}(W_i) - \widehat{\mathcal{R}}^{(i)}(W_{i+1}).$$

Proof. For notational convenience, define $g_k(W) := y_k f(x_k; W)$ and $g_k^{(i)}(W) := y_k f^{(i)}(x_k; W)$, whereby $\nabla g_k(W) = y_k \nabla f(x_k; W)$. Since ℓ is $1/4$-smooth and since, for every example (x_k, y_k), $\|\nabla f(x_k; V)\|^2 = \rho^2 \sum_{j=1}^m \|a_j \mathbb{1}[w_j^\intercal x_k \ge 0] x_k\|^2/m \le 1$, then

$$\begin{aligned}
\ell(g_k^{(i)}(W_{i+1})) &\le \ell(g_k^{(i)}(W_i)) + \ell'(g_k^{(i)}(W_i))(g_k^{(i)}(W_{i+1}) - g_k^{(i)}(W_i)) + \frac{1}{8}\left(g_k^{(i)}(W_{i+1}) - g_k^{(i)}(W_i)\right)^2 \\
&= \ell(g_k^{(i)}(W_i)) + \left\langle \ell'(g_k^{(i)}(W_i))\nabla g_k(W_i), W_{i+1} - W_i\right\rangle + \frac{1}{8}\left\langle \nabla g_k(W_i), W_{i+1} - W_i\right\rangle^2 \\
&= \ell(g_k^{(i)}(W_i)) - \eta\left\langle \ell'(g_k^{(i)}(W_i))\nabla g_k(W_i), \nabla\widehat{\mathcal{R}}(W_i)\right\rangle + \frac{1}{8}\left\langle \nabla g_k(W_i), \eta\nabla\widehat{\mathcal{R}}(W_i)\right\rangle^2 \\
&\le \ell(g_k^{(i)}(W_i)) - \eta\left\langle \ell'(g_k^{(i)}(W_i))\nabla g_k(W_i), \nabla\widehat{\mathcal{R}}(W_i)\right\rangle + \frac{\rho^2\eta^2}{8}\left\|\nabla\widehat{\mathcal{R}}(W_i)\right\|^2,
\end{aligned}$$

which after averaging over examples gives

$$\begin{aligned}\widehat{\mathcal{R}}^{(i)}(W_{i+1}) &\leq \widehat{\mathcal{R}}^{(i)}(W_i) - \frac{\eta}{n}\sum_{k=1}^{n}\left\langle \ell'(g_k^{(i)}(W_i))\nabla g_k(W_i), \nabla\widehat{\mathcal{R}}(W_i)\right\rangle + \frac{\rho^2\eta^2}{8}\left\|\nabla\widehat{\mathcal{R}}(W_i)\right\|^2 \\ &= \widehat{\mathcal{R}}^{(i)}(W_i) - \eta(1-\rho^2\eta/8)\left\|\nabla\widehat{\mathcal{R}}(W_i)\right\|^2,\end{aligned}$$

which rearranges to give the first inequality. Lastly, note if $\eta \leq 4/\rho^2$, then $\eta\left(1-\rho^2\eta/8\right) \geq \eta/2$. □

Next, the familiar regret inequality, making use of feature mappings induced by specific gradient descent iterates. Note that this inequality does not need to make any assumptions on nonlinearity and activation changes, though such effects must be controlled in the eventual application of this bound.

Lemma B.10. *For any step size $\eta \leq 4/\rho^2$, any $Z \in \mathbb{R}^{m\times d}$ and any t,*

$$\|W_t - Z\|^2 + 2\eta\sum_{i<t}\widehat{\mathcal{R}}^{(i)}(W_{i+1}) \leq \|W_0 - Z\|^2 + 2\eta\sum_{i<t}\widehat{\mathcal{R}}^{(i)}(Z).$$

Proof. As usual, using Lemma B.9,

$$\begin{aligned}\|W_{i+1} - Z\|^2 &= \|W_i - Z\|^2 - 2\eta\left\langle\nabla\widehat{\mathcal{R}}(W_i), W_i - Z\right\rangle + \eta^2\|\nabla\widehat{\mathcal{R}}(W_i)\|^2 \\ &\leq \|W_i - Z\|^2 + 2\eta\left\langle\nabla\widehat{\mathcal{R}}(W_i), Z - W_i\right\rangle + 2\eta\left(\widehat{\mathcal{R}}^{(i)}(W_i) - \widehat{\mathcal{R}}^{(i)}(W_{i+1})\right),\end{aligned}$$

where

$$\begin{aligned}\left\langle\nabla\widehat{\mathcal{R}}(W_i), Z - W_i\right\rangle &= \frac{1}{n}\sum_k \ell'(y_k f(x_k; W_i))\left\langle y_k\nabla f(x_k; W_i), Z - W_i\right\rangle \\ &= \frac{1}{n}\sum_k \ell'(y_k f(x_k; W_i))\left(y_k f^{(i)}(x_k; Z) - y_k f^{(i)}(x_k; W_i)\right) \\ &\leq \frac{1}{n}\sum_k \left(\ell(y_k f^{(i)}(x_k; Z)) - \ell(y_k f^{(i)}(x_k; W_i))\right) \\ &= \widehat{\mathcal{R}}^{(i)}(Z) - \widehat{\mathcal{R}}^{(i)}(W_i),\end{aligned}$$

together giving

$$\|W_{i+1} - Z\|^2 \leq \|W_i - Z\|^2 + 2\eta\left(\widehat{\mathcal{R}}^{(i)}(Z) - \widehat{\mathcal{R}}^{(i)}(W_{i+1})\right),$$

which after telescoping and rearranging gives the final bound. □

Lastly, the proof of Lemma 2.2, the central optimization guarantee, which immediately yields the bulk of Theorem 1.1.

Proof of Lemma 2.2. The start of this proof establishes a few inequalities used throughout. By the second part of Lemma B.7, with probability at least $1 - 3n\delta$, for any iterations (i, j) with $\|W_i - W_0\| \leq B$ and $\|W_j - W_0\| \leq B$,

$$\sup_{\|V - W_0\| \leq B} \frac{\widehat{\mathcal{R}}^{(i)}(V)}{\widehat{\mathcal{R}}^{(j)}(V)} \leq e^{\tau}. \tag{5}$$

Crucially, eq. (5) holds with $V := Z$, since $B \geq R_Z$ by definition. Additionally, by Lemma B.10, the following inequality holds *unconditionally* for every $j \leq t$:

$$\|W_j - Z\|^2 + 2\eta\sum_{i<j}\widehat{\mathcal{R}}^{(i)}(W_{i+1}) \leq \|W_0 - Z\|^2 + 2\eta\sum_{i<j}\widehat{\mathcal{R}}^{(i)}(Z). \tag{6}$$

The remainder of the proof is broken into three parts, for the three separate guarantees:

$$\|W_{\leq t} - W_0\| \leq B \qquad \text{(norm)}, \tag{7}$$

$$\widehat{\mathcal{R}}(W_{\leq t}) \leq e^{2\tau}\widehat{\mathcal{R}}^{(0)}(Z) + e^{\tau}(\rho R_Z)^2\epsilon_{\text{gd}} \qquad \text{(empirical risk)}, \tag{8}$$

$$\mathcal{R}(W_{\leq t}) \leq e^{4\tau}\mathcal{R}^{(0)}(Z) + e^{3\tau}(\rho R_Z)^2\epsilon_{\text{gd}} + e^{4\tau}\rho(B + R_Z)\tau_n \qquad \text{(risk)}. \tag{9}$$

Norm guarantee (cf. eq. (7)). There are two cases to consider: $B = R_{\mathrm{gd}}$, or $B < R_{\mathrm{gd}}$. If $B = R_{\mathrm{gd}}$, the claim follows by the definition of $W_{\leq t}$.

Now suppose $B < R_{\mathrm{gd}}$, meaning $B = 3R_Z + 2e\sqrt{\eta t \widehat{\mathcal{R}}^{(0)}(Z)}$. It will now be argued via contradiction that $\max_{i \leq t} \|W_i - W_0\| \leq B$. Assume contradictorily the claim does not hold, and let $s \leq t$ be the earliest violation. But that means the claim holds for all $i < s$, which also means, combining eq. (5) (which must hold for all $i < s$) and eq. (6) and using $\tau \leq 2$ and $\ell \geq 0$,

$$\begin{aligned}
B^2 < \|W_s - W_0\|^2 &\leq 2\|W_s - Z\|^2 + 2\|Z - W_0\|^2 \\
&\leq 2\|W_s - Z\|^2 + 4\eta \sum_{i<s} \widehat{\mathcal{R}}^{(i)}(W_{i+1}) + 2\|Z - W_0\|^2 \\
&\leq 4\|W_0 - Z\|^2 + 4\eta \sum_{i<s} \widehat{\mathcal{R}}^{(i)}(Z) \\
&\leq 4\|W_0 - Z\|^2 + 4\eta t e^2 \widehat{\mathcal{R}}^{(0)}(Z) \\
&\leq \left(2\|W_0 - Z\| + 2e\sqrt{\eta t \widehat{\mathcal{R}}^{(0)}(Z)}\right)^2 \leq B^2,
\end{aligned}$$

a contradiction.

Empirical risk guarantee (cf. eq. (8)). Now let T denote the earliest time when $\|W_i - W_0\| > 2R_Z$, or $T = \infty$ if this situation never occurs. Note that for any $i < T$,

$$\|W_i - W_0\| \leq 2R_Z \leq B,$$

and even for W_T,

$$\|W_T - W_0\| \leq \|W_{T-1} - W_0\| + \eta\|\nabla \widehat{\mathcal{R}}(W_{T-1})\| \leq 2R_Z + \eta\rho \leq B;$$

as such, eq. (5) holds for all W_i with $i \leq T$, including the edge case W_T. The remainder of the proof divides into two cases: either $T > t$ (which includes the situation $T = \infty$), or $T \leq t$.

If $T \leq t$, by the triangle inequality,

$$2\|Z - W_0\| < \|W_T - W_0\| \leq \|Z - W_T\| + \|Z - W_0\|,$$

which rearranges to give $\|Z - W_0\| < \|Z - W_T\|$, and thus, by eq. (6),

$$\begin{aligned}
\|Z - W_0\|^2 + 2\eta \sum_{i<T} e^{-\tau} \widehat{\mathcal{R}}(W_{i+1}) < \|W_T - Z\|^2 + 2\eta \sum_{i<T} \widehat{\mathcal{R}}^{(i)}(W_{i+1}) \\
\leq \|Z - W_0\|^2 + 2\eta \sum_{i<T} \widehat{\mathcal{R}}^{(i)}(Z) \\
\leq \|Z - W_0\|^2 + 2\eta \sum_{i<T} e^{\tau} \widehat{\mathcal{R}}^{(0)}(Z),
\end{aligned}$$

which after canceling from both sides and using the definition of $W_{\leq t}$,

$$\widehat{\mathcal{R}}(W_{\leq t}) \leq \min_{i<T} \widehat{\mathcal{R}}(W_i) \leq \frac{1}{T} \sum_{i<T} \widehat{\mathcal{R}}(W_i) \leq e^{2\tau} \widehat{\mathcal{R}}^{(0)}(Z),$$

establishing eq. (8) when $T \leq t$.

If $T > t$, the proof is simpler: since $\max_{i \leq t} \|W_i - W_0\| \leq 2R_Z \leq B$, then eq. (5) holds for all W_i with $i \leq t$, and thus by eq. (6) and the definition of $W_{\leq t}$,

$$\begin{aligned}
2\eta e^{-\tau} \sum_{i<t} \widehat{\mathcal{R}}(W_{i+1}) \leq 2\eta \sum_{i<t} \widehat{\mathcal{R}}^{(i)}(W_{i+1}) \leq \|W_t - Z\|^2 + 2\eta \sum_{i<t} \widehat{\mathcal{R}}^{(i)}(W_{i+1}) \\
\leq \|W_0 - Z\|^2 + 2\eta \sum_{i<t} \widehat{\mathcal{R}}^{(i)}(Z) \\
\leq \|W_0 - Z\|^2 + 2t\eta e^{\tau} \widehat{\mathcal{R}}^{(0)}(Z),
\end{aligned}$$

which after rearranging and using the definition of $W_{\leq t}$ gives

$$\widehat{\mathcal{R}}(W_{\leq t}) \leq \frac{1}{t} \sum_{i<t} \widehat{\mathcal{R}}(W_{i+1}) \leq e^{2\tau} \widehat{\mathcal{R}}^{(0)}(Z) + \frac{e^{\tau}\|W_0 - Z\|^2}{2t\eta} \leq e^{2\tau} \widehat{\mathcal{R}}^{(0)}(Z) + e^{\tau}(\rho R_Z)^2 \epsilon_{\mathrm{gd}},$$

completing the proof of eq. (8).

Risk guarantee (cf. eq. (9)). By Lemma B.8 applied once with radius B and once with radius R_Z, with probability at least $1 - 12\delta$

$$\mathcal{R}^{(0)}(W_{\leq t}) \leq \widehat{\mathcal{R}}^{(0)}(W_{\leq t}) + \rho B \tau_n, \qquad \widehat{\mathcal{R}}^{(0)}(Z) \leq \mathcal{R}^{(0)}(Z) + \rho R_Z \tau_n.$$

Moreover, by the last part of Lemma B.7 applied with radius B with probability at least $1 - 4\delta$,

$$\mathcal{R}(W_{\leq t}) \leq e^{\tau} \mathcal{R}^{(0)}(W_{\leq t}).$$

Combining all these inequalities with the empirical risk guarantee,

$$\begin{aligned}
\mathcal{R}(W_{\leq t}) &\leq e^{\tau} \mathcal{R}^{(0)}(W_{\leq t}) \\
&\leq e^{\tau} \widehat{\mathcal{R}}^{(0)}(W_{\leq t}) + e^{\tau} \rho B \tau_n \\
&\leq e^{2\tau} \widehat{\mathcal{R}}(W_{\leq t}) + e^{\tau} \rho B \tau_n \\
&\leq e^{4\tau} \widehat{\mathcal{R}}^{(0)}(Z) + e^{3\tau} (\rho R_Z)^2 \epsilon_{\text{gd}} + e^{\tau} \rho B \tau_n \\
&\leq e^{4\tau} \mathcal{R}^{(0)}(Z) + e^{3\tau} (\rho R_Z)^2 \epsilon_{\text{gd}} + \rho \left(e^{\tau} B + e^{4\tau} R_Z \right) \tau_n,
\end{aligned}$$

thus establishing eq. (9) and completing the proof. □

B.4 Approximation proofs

First, the lemma and proof that we can sample from $\overline{U}_\infty$; as the gap is over the risk, the proof uses the technique in Lemma B.3 to control all points on the sphere. This proof also makes crucial use of the $\arccos$ bound in Lemma B.4.

Lemma B.11. *Let $\overline{U}_\infty$ be given with $R := \sup_{v \in \mathbb{R}^d} \|\overline{U}_\infty(v)\|$, and suppose $m \geq \ln(emd)$. With probability at least $1 - 6\delta$,*

$$\mathcal{R}^{(0)}(\overline{U}) \leq e^{\tau} \mathcal{R}(\overline{U}_\infty), \qquad \text{where } \tau \leq 6\rho d \ln(emd^2/\delta) + \frac{20R\sqrt{d \ln(em^2 d^3/\delta)}}{m^{1/4}}.$$

Proof of Lemma B.11. Throughout this proof, the subscript will be dropped and simply $W := W_0$, with rows $(w_j^{\intercal})_{j=1}^m$.

The bound on $\mathcal{R}^{(0)}(\overline{U}) - \mathcal{R}(\overline{U}_\infty)$ follows by showing that with probability at least $1 - 6\delta$,

$$\sup_{\|x\| \leq 1} \left| f(x; \overline{U}_\infty) - f^{(0)}(x; \overline{U}) \right| \leq \tau,$$

and then as usual applying Lemma B.1 and taking an expectation to obtain a bound between $\mathcal{R}^{(0)}(\overline{U})$ and $\mathcal{R}(\overline{U}_\infty)$. Meanwhile, this intermediate bound is first established for any fixed $x \in \mathbb{R}^d$, and then general $\|x\| \leq 1$ are handled via Lemma B.3.

Fix an example $x \in \mathbb{R}^d$ and failure probability δ_0 to be determined later when Lemma B.3 is invoked. To first calculate the expected difference, note by definition of $\overline{U}$ that

$$\begin{aligned}
\mathbb{E} \left\langle \nabla f(x; W), \overline{U} - W \right\rangle &= \mathbb{E} \frac{\rho}{\sqrt{m}} \sum_{j=1}^m a_j \left\langle \overline{u}_j - w_j, x \mathbb{1}[w_j^{\intercal} x \geq 0] \right\rangle \\
&= \frac{1}{m} \sum_{j=1}^m \mathbb{E} \left\langle \overline{U}_\infty(w_j), x \mathbb{1}[w_j^{\intercal} x \geq 0] \right\rangle \\
&= f(x; \overline{U}_\infty),
\end{aligned}$$

whereas

$$\mathbb{E} \left\langle \nabla f(x; W), W \right\rangle = \mathbb{E}_a \sum_{j=1}^m a_j \mathbb{E}_{w_j} \sigma_{\text{r}}(w_j^{\intercal} x) = 0,$$

thus

$$\mathbb{E} f^{(0)}(x; \overline{U}) = \mathbb{E} \left(f^{(0)}(x; \overline{U} - W) + f^{(0)}(x; W) \right) = f(x; \overline{U}_\infty).$$

Controlling the deviations (still for this fixed x) will also consider the terms separately. The term $f^{(0)}(x;\overline{U}-W)$ will use McDiarmid's inequality; to verify the bounded differences property, consider pairs (a,W) and (a',W') which differ in only one element (a'_j,w'_j), which also defines pairs $\overline{U}$ and $\overline{U}'$ differing in just one j, meaning the vectors $\overline{u}_j$ and $\overline{u}'_j$; by Cauchy-Schwarz and the definition of R,

$$\begin{aligned}
&\left|\left\langle \nabla f(W), U-W\right\rangle - \left\langle \nabla f(W'), U'-W'\right\rangle\right| \\
&= \left|\frac{\rho}{\sqrt{m}} a_j \left\langle \overline{u}_j - w_j, x\mathbb{1}[w_j^\intercal x_j \geq 0]\right\rangle - \frac{\rho}{\sqrt{m}} a'_j \left\langle \overline{u}'_j - w'_j, x\mathbb{1}[(w'_j)^\intercal x_j \geq 0]\right\rangle\right| \\
&= \frac{1}{m}\left|a_j^2 \left\langle \overline{U}_\infty(w_j), x\mathbb{1}[w_j^\intercal x_j \geq 0]\right\rangle - (a'_j)^2 \left\langle \overline{U}_\infty(w'_j), x\mathbb{1}[(w'_j)^\intercal x_j \geq 0]\right\rangle\right| \\
&\leq \frac{2R\|x\|}{m}.
\end{aligned}$$

Thus, by McDiarmid's inequality, with probability at least $1-2\delta_0$,

$$\left|f^{(0)}(x;\overline{U}) - f(x;\overline{U}_\infty)\right| = \left|f^{(0)}(x;\overline{U}) - \mathbb{E}_{a,W} f^{(0)}(x;\overline{U})\right| \leq \sqrt{\frac{2R^2\|x\|^2\ln(1/\delta_0)}{m}}.$$

Meanwhile, the term $f^{(0)}(x;W)$ is explicitly controlled in in the first part of Lemma B.6: with probability at least $1-3\delta_0$,

$$|f^{(0)}(x;W)| \leq 2\rho\|x\|\ln(1/\delta_0).$$

Together, with probability at least $1-5\delta_0$,

$$\left|f^{(0)}(x;\overline{U}) - f(x;\overline{U}_\infty)\right| \leq 2\rho\ln(1/\delta_0) + R\sqrt{\frac{2\ln(1/\delta_0)}{m}} =: r_2.$$

Controlling the behavior for all $\|x\|\leq 1$ simultaneously now relies upon Lemma B.3, but invoked to control a single matrix, namely choosing $\mathcal{S}_0 := \{\overline{U}\}$, and radius $R_V := R/\rho \geq \|\overline{U}-W\|$. For the sake of applying Lemma B.3, define for any $V\in\mathbb{R}^{m\times d}$ the mapping

$$h_V(x) := f(x;W) - f(x;\overline{U}_\infty),$$

which has no dependence on V, and note a corresponding function $h\in\mathcal{H}$ as defined in Lemma B.3 has the form

$$h(x) = f(x;W) - f(x;\overline{U}_\infty) + \left\langle \nabla f(x;W), V-W\right\rangle = \left\langle \nabla f(x;W), V\right\rangle - f(x;\overline{U}_\infty);$$

since $\mathcal{S}_0 = \{\overline{U}\}$, we only need to check the conditions of Lemma B.3 for $V=\overline{U}$. As above, for any fixed $\|x\|\leq 1$, with probability at least $1-5\delta_0$, $|h(x)|\leq r_2$. To invoke Lemma B.3, the restricted continuity property must be established. Specifically, let $\|x-z\|\leq\epsilon$ be given, with $\epsilon>0$ determined later. Writing

$$\left|h_V(x)-h_V(z)\right| \leq \left|f(x;W)-f(z;W)\right| + \left|f(x;\overline{U}_\infty) - f(z;\overline{U}_\infty)\right|,$$

it suffices to check the restricted continuity property in both terms separately. For the first term, by Lemma B.2, with probability at least $1-\delta_0$,

$$\|W\|_2 \leq \sqrt{m}+\sqrt{d}+\sqrt{2\ln(1/\delta_0)},$$

whereby the 1-Lipschitz property of the ReLU over vectors gives

$$\left|f(x;W)-f(z;W)\right| \leq \rho\|\sigma_{\mathrm{r}}(Wx)-\sigma_{\mathrm{r}}(Wz)\| \leq \rho\|W(x-z)\| \leq \rho\left(\sqrt{m}+\sqrt{d}+\sqrt{2\ln(1/\delta_0)}\right)\epsilon.$$

For the other term, first note by a standard Gaussian calculation that

$$\begin{aligned}
\left|f(z;\overline{U}_\infty) - f(x;\overline{U}_\infty)\right| &= \left|\int \left\langle \overline{U}_\infty(v), z\mathbb{1}[v^\intercal z\geq 0] - x\mathbb{1}[v^\intercal x\geq 0]\right\rangle \mathrm{d}\mathcal{N}(v)\right| \\
&\leq R\int \left\|z\mathbb{1}[v^\intercal z\geq 0] - x\mathbb{1}[v^\intercal x\geq 0]\right\| \mathrm{d}\mathcal{N}(v) \\
&\leq R\|z-x\| \Pr_{v\sim\mathcal{N}}\left[\mathbb{1}[v^\intercal z\geq 0] = \mathbb{1}[v^\intercal x\geq 0]\right] \\
&\quad + R(\|x\|+\|z\|)\Pr_{v\sim\mathcal{N}}\left[\mathbb{1}[v^\intercal z\geq 0] \neq \mathbb{1}[v^\intercal x\geq 0]\right] \\
&\leq R\|z-x\| + R(\|x\|+\|z\|)\frac{2\arccos(\left\langle x/\|x\|, z/\|z\|\right\rangle)}{2\pi}.
\end{aligned}$$

If $\|x\| \leq 2\epsilon$, then $z \leq 3\epsilon$, and the last term can be upper bounded as $5R\epsilon$. On the other hand, if $\|x\| > 2\epsilon$, whereby $\|x\| + \|z\| \leq 2\|x\| + \epsilon \leq 3\|x\|$, then Lemma B.4 implies

$$R(\|x\| + \|z\|)\frac{2\arccos(\langle x/\|x\|, z/\|z\|\rangle)}{2\pi} \leq R(\|x\| + \|z\|)\frac{\epsilon\sqrt{8}}{\|x\|\pi} \leq 3R\epsilon,$$

Thus, by Lemma B.3 with radius $R_V := R/\rho$ and filter set $\mathcal{S}_0 := \{\overline{U}\}$ as above, and additionally choosing $\epsilon := 1/(dm)$, with overall probability at least $1 - (1 + 5(\sqrt{d}/\epsilon)^d)\delta_0$,

$$\begin{aligned}
\sup_{\|x\|\leq 1}\left|f^{(0)}(x;\overline{U}) - f(x;\overline{U}_\infty)\right| &\leq 2\rho\ln(e/\delta_0) + R\sqrt{\frac{2\ln(e/\delta_0)}{m}} \\
&\quad + \epsilon\rho\left(\sqrt{m} + \sqrt{d} + \sqrt{2\ln(e/\delta_0)}\right) + (1+5)R\epsilon \\
&\quad + 11R_V\rho\left(\frac{\ln(edm/\delta_0)}{m}\right)^{1/4} \\
&\leq 6\rho\ln(e/\delta_0) + 20R_V\rho\left(\frac{\ln(edm/\delta_0)}{m}\right)^{1/4},
\end{aligned}$$

and the final bound comes via the choice $\delta_0 := \delta/(md^2)^d$. □

The next result establishes that for any p_y, there exists a conditional probability model defined by $\overline{U}_\infty$ which is arbitrarily close, which is one of the keys to the consistency proof (cf. Corollary 2.3). As discussed briefly in Remark 1.1, this construction requires a bias term, which is simulated by replacing the input $x \in \mathbb{R}^d$ with $(x, 1)/\sqrt{2} \in \mathbb{R}^{d+1}$, and otherwise proceeding without modification.

Lemma B.12. *Suppose μ_x and p_y are Borel measurable, and μ_x is supported on $\|x\| \leq 1$. Given any $x \in \mathbb{R}^d$, let $\tilde{x} := (x, 1)/\sqrt{2} \in \mathbb{R}^{d+1}$ denote the vector obtained by appending the constant 1. Then for any $\epsilon > 0$, there exist infinite-width weights $\overline{U}_\infty : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ satisfying $R := \sup_{\tilde{v}\in\mathbb{R}^{d+1}} \overline{U}_\infty(\tilde{v}) < \infty$ and*

$$\mathcal{R}(\overline{U}_\infty) \leq \overline{\mathcal{R}} + \epsilon.$$

Proof. Throughout this proof, define $\tau := \min\{\epsilon/4, 1/2\}$.

As is standard in the theory of classification calibration [Zhang, 2004, Bartlett et al., 2006], for the logistic loss, the optimal population risk is achieved by a measurable function $\bar{f} : \mathbb{R} \to \bar{\mathbb{R}}$ which satisfies

$$\bar{f}(x) := \operatorname*{arg\,min}_{r\in\mathbb{R}\cup\pm\infty} p_y(x)\ell(r) + (1 - p_y(x))\ell(-r) = \phi^{-1}(p_y(x)) = \ln\frac{p_y(x)}{1 - p_y(x)} \qquad \mu_x\text{-a.e. } x,$$

which may take on the values $\pm\infty$. To avoid these $\pm\infty$, define a clamping of p_y as

$$p_1(x) := \max\{\tau, \min\{1 - \tau, p_y(x)\}\},$$

and clamped logits $f_1(x) := \phi^{-1}(p_1(x))$ (which now is bounded). As is again usual in the literature on classification calibration [Zhang, 2004, Bartlett et al., 2006],

$$\begin{aligned}
\mathcal{R}(f_1) - \overline{\mathcal{R}} &= \int\left(p_y(x)\ln\frac{p_y(x)}{p_1(x)} + (1 - p_y(x))\ln\frac{1 - p_y(x)}{1 - p_1(x)}\right)\mathrm{d}\mu_x(x) \\
&= \int_{p_y(x)\in[0,\tau)}\left(p_y(x)\ln\frac{p_y(x)}{\tau} + (1 - p_y(x))\ln\frac{1 - p_y(x)}{1 - \tau}\right)\mathrm{d}\mu_x(x) \\
&\quad + \int_{p_y(x)\in(1-\tau,1]}\left(p_y(x)\ln\frac{p_y(x)}{1 - \tau} + (1 - p_y(x))\ln\frac{1 - p_y(x)}{\tau}\right)\mathrm{d}\mu_x(x) \\
&\leq \frac{\tau}{1 - \tau} \leq 2\tau.
\end{aligned}$$

Since p_1 is Borel measurable (due to Borel measurability of p_y), then f_1 is Borel measurable (since ϕ^{-1} is continuous along $[\tau, 1 - \tau]$), and therefore we may apply Lusin's Theorem [Folland, 1999, Theorem 7.10]: there exists a continuous function g and a set S satisfying

$$|g| \leq |f_1| \leq \sup_x |f_1(x)| < \infty, \qquad g_{|S} = (f_1)_{|S}, \qquad \mu_x(S^c) \leq \frac{\tau}{\ell(0) + \sup_x |f_1(x)|},$$

whereby since ℓ is 1-Lipschitz,

$$\begin{aligned}\mathcal{R}(g) - \mathcal{R}(f_1) &\leq \int \mathbb{1}[x \in S^c]\ell(-yg(x))\,\mathrm{d}\mu(x,y)\\ &\leq \int \mathbb{1}[x \in S^c]\ell(|g(x)|)\,\mathrm{d}\mu_x(x)\\ &\leq \mu_x(S^c)(\ell(0) + \sup_x |g(x)|)\\ &\leq \tau.\end{aligned}$$

Since g is continuous, it is uniformly continuous over $\|x\| \leq 1$, and thus there exists a $\delta > 0$ so that the modulus of continuity $\omega_g(\delta)$ at scale δ is at most τ, meaning

$$\sup_{\|x-x'\|\leq\delta} |g(x) - g(x')| \leq \omega_g(\delta) \leq \tau.$$

By results in neural network universal approximation [Ji et al., 2020b, Theorem 4.3], there exists infinite-width weights $\overline{U}_\infty : \mathbb{R}^{d+1} \to \mathbb{R}^{d+1}$ satisfying $R := \sup_{\tilde{x}} \|\overline{U}_\infty(\tilde{x})\| < \infty$ and

$$\sup_{\|x\|\leq 1} \left| f(\tilde{x}; \overline{U}_\infty) - g(x) \right| \leq \omega_g(\delta) \leq \tau,$$

which again by the 1-Lipschitz property of ℓ means $\mathcal{R}(\overline{U}_\infty) - \mathcal{R}(g) \leq \tau$. Combining all these pieces,

$$\mathcal{R}(\overline{U}_\infty) - \overline{\mathcal{R}} = \left[\mathcal{R}(\overline{U}_\infty) - \mathcal{R}(g)\right] + \left[\mathcal{R}(g) - \mathcal{R}(f_1)\right] + \left[\mathcal{R}(f_1) - \overline{\mathcal{R}}\right] \leq \tau + \tau + 2\tau \leq \epsilon,$$

as desired. □

B.5 Proofs of main results: Theorem 1.1 and Corollary 2.3

The proof of Theorem 1.1 and a precise restatement are as follows. This restatement has fully explicit constants, and is invoked in the proof of Corollary 2.3 to ease sanity-checking.

Theorem B.13 (Refined restatement of Theorem 1.1)**.** *Let temperature $\rho > 0$ and reference model $\overline{U}_\infty$ be given with $R := \max\{4, \rho, \sup_v \|\overline{U}_\infty(v)\|\} < \infty$, and define a corresponding conditional model $\phi_\infty(x) := \phi(f(x; \overline{U}_\infty))$. Let optimization accuracy ϵ_{gd} and radius $R_{\text{gd}} \geq R/\rho$ be given, define effective radius $B := \min\left\{R_{\text{gd}}, \frac{3R}{\rho} + \frac{4e}{\rho}\sqrt{t}\sqrt{e^{\tau_0}\mathcal{R}(\overline{U}_\infty) + R\tau_n}\right\}$, where generalization error τ_n and additionally linearization error τ_1 and sampling error τ_0 are defined as*

$$\begin{aligned}\tau_n &:= \frac{80\left(d\ln(em^2d^3/\delta)\right)^{3/2}}{\sqrt{n}},\\ \tau_1 &:= \frac{100\rho B^{4/3}\sqrt{d\ln(enm^2d^3/\delta)}}{m^{1/6}},\\ \tau_0 &:= 6\rho d\ln(emd^2/\delta) + \frac{20R\sqrt{d\ln(em^2d^3/\delta)}}{m^{1/4}},\end{aligned}$$

where it is assumed $\tau_1 \leq 2$ and $m \geq \ln(emd)$. Choose step size $\eta := 4/\rho^2$, and run gradient descent for $t := 1/(8\epsilon_{\text{gd}})$ iterations, selecting iterate $W_{\leq t} := \arg\min\{\widehat{\mathcal{R}}(W_i) : i \leq t, \|W_i - W_0\| \leq R_{\text{gd}}\}$ with simultaneously small norm and empirical risk. Then, with probability at least $1 - 25\delta$,

$$\begin{aligned}&\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} && \textit{(logistic error)}\\ \leq\ &\mathcal{K}_{\text{bin}}(p_y, \phi_\infty) + \left(e^{\tau_1+\tau_0} - 1\right)\mathcal{R}(\overline{U}_\infty) && \textit{(reference model error)}\\ +\ &e^{\tau_1}R^2\epsilon_{\text{gd}} && \textit{(optimization error)}\\ +\ &e^{\tau_1}(\rho B + R)\tau_n && \textit{(generalization error)},\end{aligned}$$

where the classification and calibration errors satisfy

$$\begin{aligned}&\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} && \textit{(logistic error)}\\ \geq\ &2\int \left(\phi(f(x; W_{\leq t})) - p_y\right)^2 \mathrm{d}\mu_x(x) && \textit{(calibration error)}\\ \geq\ &\frac{1}{2}\left(\mathcal{R}_{\text{z}}(W_{\leq t}) - \overline{\mathcal{R}}_{\text{z}}\right)^2 && \textit{(classification error)}.\end{aligned}$$

Lastly, for any $\epsilon > 0$, there exists $\overline{U}_\infty^{(\epsilon)}$ with $\sup_v \|\overline{U}_\infty^{(\epsilon)}(v)\| < \infty$ and whose conditional model $\phi_\infty^{(\epsilon)}(x) := \phi(f((x,1)/\sqrt{2}; \overline{U}_\infty^{(\epsilon)}))$ satisfies $\mathcal{K}_{\mathsf{bin}}(p_y, \phi_\infty^{(\epsilon)}) \leq \epsilon$.

Proof of Theorem 1.1 and simultaneously Theorem B.13. This proof focuses on the first inequality, upper bounding $\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}}$; for the other two statements, the chain of inequalities with other error metrics are from Lemma B.1, and the approximation of arbitrary Borel measurable p_y is from Lemma B.12. (The only difference between Theorem B.13 here and Theorem 1.1 in the body is that the "$\widetilde{\mathcal{O}}$" hides constants and $\ln(m)$ and $\ln(d)$ (but not $\ln(n)$).

Returning to the first inequality, let $\overline{U}$ be the canonical sample of $\overline{U}_\infty$ as in eq. (1), where $\|\overline{U} - W_0\| \leq R/\rho$ by construction. By Lemma B.11, with probability at least $1 - 6\delta$, then $\mathcal{R}^{(0)}(\overline{U}) \leq e^{\tau_0}\mathcal{R}(\overline{U}_\infty)$, where τ_0 is as in the statement (cf. Theorem B.13).

Next instantiate Lemma 2.2 with reference matrix $Z = \overline{U}$ and $R_Z := R/\rho$, whereby the definition of R gives $R_Z \geq \{1, \eta\rho, \|\overline{U} - W_0\|\}$ as needed; as such, ignoring an additional failure probability at most 19δ, setting $\tau := \tau_1/4$ in the invocation, and lastly subtracting $\overline{\mathcal{R}}$ from both sides,

$$\begin{aligned}\mathcal{R}(W_{\leq t}) - \overline{\mathcal{R}} &\leq e^{\tau_1}\mathcal{R}^{(0)}(\overline{U}) + e^{\tau_1}(\rho R_Z)^2\epsilon_{\mathsf{gd}} + e^{\tau_1}(\rho B + \rho R_Z)\tau_n - \overline{\mathcal{R}} \\ &\leq \left(e^{\tau_1+\tau_0} - 1\right)\mathcal{R}(\overline{U}_\infty) + \mathcal{K}_{\mathsf{bin}}(p_y, \phi_\infty) + e^{\tau_1}R^2\epsilon_{\mathsf{gd}} + e^{\tau_1}(\rho B + R)\tau_n.\end{aligned}$$

This invocation of Lemma 2.2 also guarantees $\widehat{\mathcal{R}}^{(0)}(\overline{U}) \leq \mathcal{R}^{(0)}(\overline{U}) + R\tau_n$ which together with the earlier inequality $\mathcal{R}^{(0)}(\overline{U}) \leq e^{\tau_0}\mathcal{R}(\overline{U}_\infty)$ provides the form of B used in the statement (this B upper bounds the one defined in Lemma 2.2, which is fine since it only relaxes the guarantees provided there). □

Making use of Theorem B.13, the proof of the consistency statement, Corollary 2.3, is as follows. Note that we are always working with bias-augmented inputs within this statement and its proof; e.g., $\widehat{W}_n \in \mathbb{R}^{m^{(n)} \times (d+1)}$.

Proof of Corollary 2.3. Let $\epsilon > 0$ be arbitrary, and define the event

$$E_n := \left[\mathcal{R}(\widehat{W}_n) \geq \overline{\mathcal{R}} + \epsilon\right].$$

Following a standard scheme for consistency proofs [Schapire and Freund, 2012, Corollary 12.3], it suffices, thanks to the Borel-Cantelli lemma, to prove

$$\sum_{n \geq 1} \Pr[E_n] < \infty; \tag{10}$$

that is to say, by the Borel-Cantelli lemma, eq. (10) implies $\limsup_{n\to\infty} \mathcal{R}(\widehat{W}_n) - \overline{\mathcal{R}} \leq \epsilon$ almost surely, and since $\mathcal{R}(\widehat{W}_n) \geq \overline{\mathcal{R}}$ and since $\epsilon > 0$ was arbitrary, it follows that $\mathcal{R}(\widehat{W}_n) \to \overline{\mathcal{R}}$ almost surely. Moreover, by Lemma B.1, for each n there are the inequalities

$$\frac{1}{2}\left(\mathcal{R}_{\mathrm{z}}(\widehat{W}_n) - \overline{\mathcal{R}}_{\mathrm{z}}\right)^2 \leq 2\int (\widehat{\phi}_n(x) - p_y(x))^2 \,\mathrm{d}\mu_x(x) \leq \mathcal{R}(\widehat{W}_n) - \overline{\mathcal{R}},$$

thus $\mathcal{R}(\widehat{W}_n) \to \overline{\mathcal{R}}$ also implies $\widehat{\phi}_n \to p_y$ in $L_2(\mu_x)$ almost surely, and $\mathcal{R}_{\mathrm{z}}(\widehat{W}_n) \to \overline{\mathcal{R}}_{\mathrm{z}}$ almost surely.

To establish eq. (10), first use the last part of Theorem B.13 to fix a $\overline{U}_\infty$ with $\mathcal{K}_{\mathsf{bin}}(p_y, \widehat{\phi}_n) \leq \epsilon/2$, and define $R := \sup_v \|\overline{U}_\infty(v)\| < \infty$. To bound $\Pr[E_n]$, instantiate Theorem B.13 for every n with reference model $\overline{U}_\infty$ and corresponding $R < \infty$, and failure probability $\delta^{(n)} := 1/n^2$, and optimization radius $R_{\mathsf{gd}} = \infty$, meaning a corresponding effective radius given by Theorem B.13 as

$$B^{(n)} = \frac{1}{\rho^{(n)}}\left(3R + 4e\sqrt{t^{(n)}}\sqrt{e^{\tau_0^{(n)}}\mathcal{R}(\overline{U}_\infty) + R\tau_n}\right).$$

Inspecting all the terms in Theorem B.13, it will now be argued that while the term $\mathcal{K}_{\mathsf{bin}}(p_y, \widehat{\phi}_n)$ stays level and is at most $\epsilon/2$ independent of n, all other terms go to 0. Returning to $B^{(n)}$, since

$\tau_n = \widetilde{\mathcal{O}}(1/\sqrt{n})$ and $\tau_0^{(n)} \to 0$ (which will be shown later), then $B^{(n)} = \widetilde{\mathcal{O}}(\sqrt{t^{(n)}}/\rho^{(n)})$, whereby

$$\tau_1^{(n)} = \widetilde{\mathcal{O}}\left(\frac{\rho^{(n)}(B^{(n)})^{4/3}}{(m^{(n)})^{1/6}}\right) = \widetilde{\mathcal{O}}\left(\frac{(t^{(n)})^{2/3}}{(m^{(n)})^{1/6}(\rho^{(n)})^{1/3}}\right)$$
$$= \widetilde{\mathcal{O}}\left(\frac{(t^{(n)})^{2/3}}{(m^{(n)})^{1/8}}\right) = \widetilde{\mathcal{O}}\left(\frac{n^{\frac{2}{3}(1-\xi)}}{n^{\frac{5}{3}(1-\xi)}}\right) = \widetilde{\mathcal{O}}\left(n^{\xi-1}\right) \to 0.$$

Next,

$$\tau_0^{(n)} = \widetilde{\mathcal{O}}\left(\rho^{(n)} + \frac{1}{(m^{(n)})^{1/4}}\right) = \widetilde{\mathcal{O}}\left(n^{\frac{5}{3}(\xi-1)} + n^{\frac{10}{3}(\xi-1)}\right) \to 0,$$

which together with the asymptotics of $\tau_1^{(n)}$ gives $\exp(\tau_0^{(n)} + \tau_1^{(n)}) - 1 \to 0$ and $\exp(\tau_1^{(n)})R^2\epsilon_{\text{gd}}^{(n)} \to 0$. The final term to consider is

$$\exp(\tau_1^{(n)})\rho^{(n)}B^{(n)}\tau_n = \widetilde{\mathcal{O}}\left(\sqrt{\frac{t^{(n)}}{n}}\right) = \widetilde{\mathcal{O}}(n^{-\xi/2}) \to 0.$$

As such, all terms go to zero with n (excepting $\mathcal{K}_{\text{bin}}(p_y, \widehat{\phi}_n) \le \epsilon/2$, which is fine), and there exists N_0 so that for all $n > N_0$, all conditions of the bound are met, and with the exclusion of a failure probability of $\delta^{(n)}$, the bound implies $\mathcal{R}(\widehat{W}_n) < \overline{\mathcal{R}} + \epsilon$. Thus $n \ge N_0$ implies $\Pr[E_n] \le \delta^{(n)} = 1/n^2$, and

$$\sum_{n \ge 1} \Pr[E_n] \le \sum_{n \le N_0} 1 + \sum_{n > N_0} \frac{1}{n^2} \le N_0 + \frac{\pi^2}{6} < \infty,$$

which establishes eq. (10) and completes the proof. □

C Proof of Proposition 1.2

Proposition 1.2 is a consequence of the following more refined statement, which also suggests the method of proof, and is consistent with Figure 2.

Lemma C.1. *Suppose marginal distribution μ_x is continuous and compactly supported on $[0,1]$, p_y is continuous, and that either $\mu_x(p_y^{-1}((0,1/2))) > 0$ or $\mu_x(p_y^{-1}((1/2,1)) > 0$, meaning p_y is outside $\{0, 1/2, 1\}$ on a set which has positive measure according to μ_x.*

Then there exists a constant $c \in (0, 1/4)$ (depending only on μ_x and p_y) so that with probability at least $1 - 7\delta$ over the draw of $((x_i, y_i))_{i=1}^n$ with $n \ge \ln(1/\delta)/c$, there exists an interval $I \subseteq [0,1]$, and a subset of pairs of indices indices $S \subseteq [m]^2$ satisfying the following properties.

1. *Either $p_y \in [c, 1/2 - c]$ everywhere on I, or $p_y \in [1/2 + c, 1 - c]$ everywhere on I; henceforth let $\hat{y} := \text{sgn}(p_y - 1/2)$ designate the correct (Bayes) prediction over I.*
2. *If $(i,k) \in S$, then $x_i < x_k = \min\{x_s : x_s \ge x_i\}$, meaning x_k is the first point to the right of x_i, and moreover the corresponding labels $y_i = y_k = -\hat{y}$ agree with each other but are incorrect.*
3. *For any local interpolation rule $f \in \mathcal{F}_n$ (cf. Proposition 1.2),*
$$\mathcal{R}_{\text{z}}(f) \ge \overline{\mathcal{R}}_{\text{z}} + c.$$

Proof of Lemma C.1 (and simultaneously Proposition 1.2). Consider any point x where $p_y(x) \notin \{0, 1/2, 1\}$ and $\mu_x > 0$; such a point must exist by the assumptions. Define $\hat{y} := \text{sgn}(p_y(x) - 1/2)$ and $c_1 := \min\{p_y(x)/2, |p_y(x) - 1/2|/2, (1 - p_y(x))/2\}$, where $c_1 \in (0, 1/4)$ by construction. Since p_y and μ_x are continuous, then there must exist some (potentially tiny) closed interval I containing x so that $\text{sgn}(p_y(x) - 1/2) = \hat{y}$, and for any $x' \in I$, both $\mu_x(x') > 0$ and $p_{x'} \in (c_1, 1/2 - c_1) \cup (1/2 + c_1, 1 - c_1)$.

To simplify the rest of the proof, suppose $\hat{y} = -1$; the other case is symmetric, but as in the preceding paragraph, handling both cases simultaneously adds significant notational overhead.

Let S denote all adjacent pairs of points in I where $(x_i, x_k) \in S$ means $x_i < x_k = \min\{x_s : x_s > x_i\}$ and $y_i = y_k = -\hat{y}$. With this choice, all that remains to be shown is the third item, the lower

bound on the risk.To show this, it suffices to show that a constant fraction of μ_x's probability mass is contained between these pairs, meaning

$$\mu_x\left(\cup_{(i,k)\in S}\mu([x_i,x_k])\right) \geq c_2 > 0,$$

where crucially c_2 is independent of n. To see that this suffices to establish the third property, suppose that $f:\mathbb{R}\to\mathbb{R}$ satisfies the required condition, meaning $f(x)\hat{y}<0$ for $x\in\cup_{(i,k)\in S}\mu([x_i,x_k])$; then by a standard calculation against the Bayes risk [Devroye et al., 1996],

$$\begin{aligned}
\mathcal{R}_z(f)-\overline{\mathcal{R}}_z &= \int |1-2p_y(x)|\mathbb{1}\left[\operatorname{sgn}(f)\neq\operatorname{sgn}(p_y(x)-1/2)\right]\mathrm{d}\mu_x(x)\\
&\geq \int |1-2p_y(x)|\mathbb{1}\left[x\in\cup_{(i,k)\in S}[x_i,x_k]\right]\mathrm{d}\mu_x(x)\\
&\geq 2c_1\mu_x\left(\cup_{(i,k)\in S}[x_i,x_k]\right)\\
&= 2c_1c_2,
\end{aligned}$$

and the final statement and all properties are satisfied if we pick $c\in\left(0,\min\{c_1,c_2,2c_1c_2\}\right]$.

As such, it remains to provide a lower bound on c_2 which is independent of n, which will follow a series of simplifications as follows.

The first step is to lower bound the cardinality of S. The expected number of points in I is $n\mu_x(I)$, and if $n\geq 32\ln(1/\delta)/\mu_x(I)$, then by a multiplicative Chernoff bound [Blum et al., 2020, Theorem 12.6], with probability at least $1-3\delta$,

$$\left|\left\{i\in[m]:x_i\in I\right\}\right| \geq \frac{n\mu_x(I)}{2}.$$

and thus the number of consecutive pairs in I is at least $n\mu_x(I)/2-1\geq n\mu_x(I)/4$.

Since these pairs may share endpoints, consider the set of at least $n\mu_x(I)/8$ pairs that share no points. Since the draw of y is independent of x, for each of these consecutive pairs, the probability that both labels are wrong is at least $(1-c_1)^2$ (and is independent of other pairs), meaning the expected number of such points is at least $n\mu_x(I)(1-c_1)^2/8$; as such, if $n\geq 256\ln(1/\delta)/(\mu_x(I)(1-c_1)^2)$, by another multiplicative Chernoff bound, with probability at least $1-3\delta$, the number of pairs with agreeing but incorrect labels is at least $n\mu_x(I)(1-c_1)^2/16$. Let S_0 denote this set of pairs; by construction, its cardinality also lower bounds that of S.

It remains to show that the union of the convex hulls of these pairs of points has a significant fraction of total probability mass.

For any sample $(x_1,\ldots,x_n)$, let $(x_{(1)},\ldots,x_{(n)})$ be the sample in sorted order, meaning $x_{(1)}<x_{(2)}<\cdots<x_{(n)}$ (strict inequalities almost surely since μ_x is continuous). Define a distance Δ and function F of the sample as

$$\begin{aligned}
\Delta &:= \frac{\mu_x(I)(1-c_1)^2}{256n},\\
F(x_1,\ldots,x_n) &:= \left|\left\{i\in[m-1]:\mu([x_{(i)},x_{(i+1)}])<\Delta\right\}\right|;
\end{aligned}$$

that is to say, F measures the number of consecutive pairs whose convex hulls have probability mass strictly less than Δ. As will be established momentarily, F satisfies the bounded differences property with a constant 2, meaning for any two samples $(x_1,\ldots,x_n)$ and $(x'_1,\ldots,x'_n)$ that differ only in a single example $x_i\neq x'_i$,

$$\left|F(x_1,\ldots,x_n)-F(x'_1,\ldots,x'_n)\right|\leq 2.$$

To argue this, suppose the disagreeing example x_i occupies position j after sorting, meaning $x_i=x_{(j)}$, and consider adjusting one sample to the other by renaming this point to x'_i, removing it from its current location, and moving it to its final location.

- First we remove x'_i from the interval $(x_{(j-1)},x_{(j+1)})$. If neither $(x_{(j-1)},x'_i)$ nor $(x'_i,x_{(j+1)})$ counts towards F, then neither will $(x_{(j-1)},x_{(j+1)})$, so F remains unchanged. If exactly

one of $(x_{(j-1)}, x_i')$ and $(x_i', x_{(j+1)})$ counts towards F, then $(x_{(j-1)}, x_{(j+1)})$ does not count towards F, so F decreases by 1. If both $(x_{(j-1)}, x_i')$ and $(x_i', x_{(j+1)})$ counts towards F, then $(x_{(j-1)}, x_{(j+1)})$ may or may not count towards F, so F decreases by 1 or 2. So this operation changes F by any of $\{-2, -1, 0\}$.

- Then we insert x_i' into a new interval. The range of possible changes to F is the exact opposite as removing it from an interval, so this leads to a change by any of $\{+2, +1, 0\}$; together the difference in F is within $[-2, +2]$.

As such, by McDiarmid's inequality, with probability at least $1 - \delta$,

$$F(x_1, \dots, x_n) \le \mathbb{E}F(x_1, \dots, x_n) + \sqrt{2n \ln(1/\delta)}.$$

Upper bounding $\mathbb{E}F(x_1, \dots, x_n)$ can now be performed in a coarse way as follows. Partition the support of μ_x, $[0, 1]$, into two systems of intervals, $\mathcal{I}$ and $\mathcal{J}$, as follows. $\mathcal{I}$ simply contains the $\lceil 1/(2\Delta) \rceil$ consecutive intervals of mass 2Δ (except for the last, which may have less mass); meanwhile, $\mathcal{J}$ contains a first initial interval of mass Δ, and then intervals of mass 2Δ until a final interval of mass at most 2Δ. Due to this staggered behavior, if some pair $(x_{(i)}, x_{(i+1)})$ has $\mu_x((x_{(i)}, x_{(i+1)})) < \Delta$, then the pair must appear in a single interval in either $\mathcal{I}$ or $\mathcal{J}$ (the staggering avoids boundary issues). Now consider the creation of the full data sample by sampling the data points one by one, and the resulting effect on these bins; the goal is to upper bound the number of times a point is inserted into an occupied bin, as this upper bounds the number of consecutive pairs of points within some bin, which in turn upper bounds F. After inserting the ith point (twice), let A_i denote the number of occupied bins, and B_i the number of times a point was inserted into an occupied bin; necessarily, $A_i = 2i - B_i$ (the factor two coming from simultaneous throws to $\mathcal{I}$ and $\mathcal{J}$). The probability of landing in an occupied bin (and thus increasing B_i) is at most $A_i(2\Delta) = (2i - B_i)(2\Delta)$. By linearity of expectation,

$$\begin{aligned}\mathbb{E}F \le \mathbb{E}B_n &\le \sum_{i=1}^{n-1} 2\mathbb{E}\mathbb{1}\left[x_{i+1} \text{ lands in an occupied bin }\right] \\ &\le 4\Delta \sum_{i=1}^{n-1} (2i - \mathbb{E}B_i) \le 4\Delta(n-1)n \le \frac{n\mu_x(I)(1-c_1)^2}{64}.\end{aligned}$$

Together, supposing that $n \ge 8192 \ln(1/\delta)/(\mu_x(I)^2(1-c_1)^4)$, it follows that with probability at least $1 - \delta$,

$$F(x_1, \dots, x_n) \le \frac{n\mu_x(I)(1-c_1)^2}{64} + \sqrt{2n\ln(1/\delta)} \le \frac{n\mu_x(I)(1-c_1)^2}{32}.$$

To finish the proof, since the preceding quantity is less than half the cardinality of S_0, we are guaranteed that at least half the pairs in S_0 have $\mu_x((x_i, x_k)) \ge \Delta$; letting S_1 denote this half, then

$$\begin{aligned}\mu_x(\cup_{i,k\in S}[x_i, x_k]) &\ge \sum_{(i,k)\in S_1} \mu_x([x_i, x_k]) \\ &\ge |S_1|\Delta \ge \frac{n\mu_x(I)(1-c_1)^2}{32} \cdot \Delta \ge \frac{\mu_x(I)^2(1-c_1)^4}{8192} =: c_3.\end{aligned}$$

It only remains to determine the final value of the constant c. By the preceding calculation and the comments near the start of the proof establishing that $c \in (0, \min\{c_1, c_2, 2c_1c_2\}]$ suffices, the quantity c_3 here is indeed a lower bound on c_2, and thus, defining $c_4 := \min\{c_1, c_3, 2c_1c_3\}$, it suffices to require $c \in (0, c_4]$. On the other hand, inspecting all the necessary lower bounds on n throughout the proof, the maximum across all of them is that we need $n \ge \ln(1/\delta)/c_3$. As such, all properties are satisfied if we take $c := c_4 > 0$ as our final constant, which depends only on μ_x and p_y (but not on n) as promised. □