Statistical Learning under Heterogenous Distribution Shift

Max Simchowitz [*1] **Anurag Ajay** [*1] **Pulkit Agrawal** [1] **Akshay Krishnamurthy** [2]

Abstract

This paper studies the prediction of a target $\mathbf{z}$ from a pair of random variables $(\mathbf{x}, \mathbf{y})$, where the ground-truth predictor is additive $\mathbb{E}[\mathbf{z} \mid \mathbf{x}, \mathbf{y}] = f_\star(\mathbf{x}) + g_\star(\mathbf{y})$. We study the performance of empirical risk minimization (ERM) over functions $f + g$, $f \in \mathcal{F}$ and $g \in \mathcal{G}$, fit on a given training distribution, but evaluated on a test distribution which exhibits covariate shift. We show that, when the class $\mathcal{F}$ is "simpler" than $\mathcal{G}$ (measured, e.g., in terms of its metric entropy), our predictor is more resilient to *heterogenous covariate shifts* in which the shift in $\mathbf{x}$ is much greater than that in $\mathbf{y}$. These results rely on a novel Hölder style inequality for the Dudley integral which may be of independent interest. Moreover, we corroborate our theoretical findings with experiments demonstrating improved resilience to shifts in "simpler" features across numerous domains.

1. Introduction

Modern machine learning systems are routinely deployed under distribution shift (Taori et al., 2020; Koh et al., 2021). However, statistical learning theory has primarily focused on studying the generalization error in the situation where the test and the training distributions are identical (Bartlett and Mendelson, 2002; Vapnik, 2006). In the setting of *covariate shift*—where only features/covariates change between training and testing, but the target function remains fixed—guarantees from statistical learning theory can be applied via a reweighting argument, leading to the classical bound involving density ratios depicted in Eq. (3.2). However, this approach may be overly pessimistic and may not account for the relative differences in performance degradation between different distribution shift settings.

Well-specified linear regression is perhaps the simplest setting that admits favorable distribution shift behavior (Lei et al., 2021). Here, out-of-distribution generalization is controlled by the alignment between the second moment matrices of the training and test distribution, rather than the significantly worse density ratios. Beyond the linear setting, ML models including neural networks often suffer from *spurious correlation* (c.f., Arjovsky et al., 2019), where the model exploits correlations in the training distribution to learn an accurate-but-incorrect predictor that fails to generalize to a de-correlated distribution. Though this phenomenon and other related ones are well-documented experimentally, a general theory of distribution shift—particularly one that explains the behavior of deep learning models in practice—has remained undeveloped.

A useful theory of distribution shift should make predictions as to which shifts a learned model is most sensitive to in possible test environments, given properties of the model which can be evaluated from training data (Xiao et al., 2020; Koh et al., 2021; Rahimian and Mehrotra, 2019). We illustrate this point with the following example.

Example 1.1. Consider a quadruped carrying different payloads across multiple terrains, with a policy trained via reinforcement learning. Should one expect more degradation in performance with new shapes or sizes of payloads? Or should a policy suffer more from novel terrains? If our policy requires camera inputs, should we expect that changes in lighting conditions or times of day have more of an effect? Or if the policy relies on tactile sensation, should we expect changes in weather (e.g., rain on the tactile sensors) to present more of an obstacle?

Contributions. This paper gestures towards a richer theory of generalization under covariate shift; one that makes such actionable predictions about the relative resilience of a model to the kinds of multifarious shifts illustrated in Example 1.1. More specifically, we highlight a setting we call *heterogenous covariate shift*, where the distribution of one feature shifts more than another.

Theoretically, we study supervised prediction from a pair of (possibly non-independent) random variables $(\mathbf{x}, \mathbf{y})$. We think of $\mathbf{x}$ as corresponding to "simple features" and $\mathbf{y}$ to more complex ones. We show greater resilience to heterogenous distribution shifts in which the shift in the marginal of $\mathbf{y}$ is significantly smaller than that of the joint

[*]Equal contribution [1]CSAIL, Massachusetts Institute of Technology [2]Microsoft Research, New York City. Correspondence to: Max Simchowitz <msimchow@csail.mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

distribution. Specifically, our analysis restricts its attention to regression functions which decompose additively as $f(\mathbf{x}) + g(\mathbf{y})$. We show that empirical risk minimization (ERM) over functions of the form $f(\mathbf{x}) + g(\mathbf{y})$ leads to much more favorable generalization guarantees than those obtained via the naïve covariate shift bound. In the most favorable setting, we obtain a test error bound that scales only with the covariate shift in the marginal of the "complex feature" $\mathbf{y}$, so that even though spurious correlations between $\mathbf{x}$ and $\mathbf{y}$ are present, they play no role in the generalization performance of the ERM.

While limited, the additive framework proposes a useful metric to evaluate relative complexity of the features: the richness of their associated function classes. This suggests a more general hypothesis that can be formulated *without the additivity assumption:* we can determine resilience to shifts in a given feature by evaluating the "complexity" of a model's dependence on that feature. Using in-distribution generalization as a proxy for model complexity, we find that deep learning models are *consistenly more resilient to shifts in simpler features than they are to shifts in complex features*; this finding holds across a range of tasks, including synthetic settings, computer vision benchmarks, and imitation learning. We hope that, taken together, our theoretical and experimental results initiate a further dialogue between the field of statistical learning theory and the study of distribution shift in machine learning more broadly.

Proof Techniques. The technical challenge to obtaining favorable distribution shift is correlation between $\mathbf{x}$ and $\mathbf{y}$, which, among other things, leads to unidentifiability of the generalizing predictor. We show that when $\mathcal{G}$ is sufficiently expressive, the simple predictor f can be learned, up to a bias arising from identifiability, at a rate that exhibits a lower order dependence on $\mathcal{G}$. Although this predictor *is* affected by distribution shifts in $\mathbf{x}$, the low complexity of the function class $\mathcal{F}$ and the lower order dependence on $\mathcal{G}$ implies that the impact on the overall performance is rather small. Then ERM can learn a g that corrects for the bias in f and is unaffected by distribution shifts in $\mathbf{x}$. The core technical result for this argument is the generalization bound for $\mathcal{F}$ which disentangles the correlations between $\mathbf{x}$ and $\mathbf{y}$; this result relies, among other things, on a novel Hölder-style inequality for the Dudley integral of products of function classes, which may be of independent interest.

Related Work. Our results and techniques are very much in the spirit of classical statistical learning theory (Bartlett et al., 2005; Bousquet and Elisseeff, 2002; Bartlett and Mendelson, 2002; Vapnik, 2006), but also have the flavor of more recent work on orthogonal/double machine learning (Chernozhukov et al., 2017; Foster and Syrgkanis, 2019; Mackey et al., 2018). In that parlance, we can view g as a nuisance parameter for estimating f and our results show similar (but not quite matching) recovery guarantees without explicit double-training interventions. We discuss comparisons to orthogonal ML in the sequel. Resilience to distribution shift has received considerable attention in recent years (Miller et al., 2021; Taori et al., 2020; Santurkar et al., 2020; Koh et al., 2021; Zhou et al., 2022), with the vast majority of the work being empirical. While the present work focuses on studying vanilla empirical risk minimization, there have been many methods produced to explicity tackle distribution shift including CORAL (Sun and Saenko, 2016), IRM (Arjovsky et al., 2019), and distributionally robust optimization, the latter having seen recent advances on both empirical and theoretical fronts (Schmidt et al., 2018; Rahimian and Mehrotra, 2019; Sinha et al., 2018). Though the statistical properties of distribution shift under empirical risk minimization has garnered substantially less attention, recent work has given precise characterizations of the effects of covariate shift for certain specific function classes, notably kernels (Ma et al., 2022) and Hölder smooth classes (Pathak et al., 2022). Our work complements these by considering structural situations in which interesting generalization phenomena arise for arbitrary function classes. Lastly, (Dong and Ma, 2023) establish Laplacian-like connectivity conditions under which test-error of additive predictors $f(\mathbf{x}) + g(\mathbf{x})$ (as in this work) can be bounded in terms of train-error, focusing on (a) situations where the marginals over $\mathbf{x}, \mathbf{y}$ between test- and train-distibutions coincide but joint distributions differ and (b) discrete- Gaussian-distributed features. By contrast, our work allows for changes in both joint and marginal distributions (albeit with cruder measures of shift), general feature distributions, and exposes statistical phenomena not addressed by the former work.

2. Theoretical Setup

We study the prediction of a scalar $\mathbf{z} \in \mathbb{R}$ from two covariates $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$ under distribution shift. We postulate a pair of testing and training environments denoted $\mathsf{e} \in \{\mathsf{test}, \mathsf{train}\}$, each of which index laws $\mathbb{P}_\mathsf{e}$ over $(\mathbf{x}, \mathbf{y}, \mathbf{z})$, and whose expectation operators are denoted by $\mathbb{E}_\mathsf{e}$. We assume the environments do not differ in the Bayes regression function, i.e., they exhibit only *covariate shift*:

Assumption 2.1 (Covariate Shift)**.** We assume that, for all $\mathbf{x}, \mathbf{y}$, $\mathbb{E}_{\mathsf{train}}[\mathbf{z} \mid \mathbf{x}, \mathbf{y}] = \mathbb{E}_{\mathsf{test}}[\mathbf{z} \mid \mathbf{x}, \mathbf{y}]$.

Next, we assume that the we have access to a class of functions that capture the conditional expectations $\mathbb{E}_{\mathsf{train}}[\mathbf{z} \mid \mathbf{x}, \mathbf{y}]$ via additive structure. Specifically, we assume access to classes $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ and $\mathcal{G} : \mathcal{Y} \to \mathbb{R}$ for which $(x, y) \mapsto \mathbb{E}_{\mathsf{train}}[\mathbf{z} \mid \mathbf{x} = x, \mathbf{y} = x] \in \mathcal{F} + \mathcal{G}$. This is typically referred to as being realizable or well-specified.

Assumption 2.2 (Additive well-specification)**.** For some

$f_\star \in \mathcal{F}$ and $g_\star \in \mathcal{G}$, it holds thats

$$\mathbb{P}_{\mathsf{train}}[\mathbf{z} \mid \mathbf{x} = x, \mathbf{y} = x] \sim \mathcal{N}(f_\star(x) + g_\star(y), \sigma^2) \quad (2.1)$$

Via universality of Gaussian processes, our results can be extended to general subgaussian noise. Since the model is well-specified, a natural performance measure of a predictor (f, g) is its excess square-loss risk, denoted $\mathcal{R}_{\mathsf{e}}(f, g)$:

$$\mathcal{R}_{\mathsf{e}}(f, g) := \mathbb{E}_{\mathsf{e}}((f - f_\star)(\mathbf{x}) + (g - g_\star)(\mathbf{y}))^2.$$

Empirical Risk Minimization. We study the excess risk under $\mathbb{P}_{\mathsf{test}}$ of square-loss empirical risk minimizers, or ERMs, for $\mathbb{P}_{\mathsf{train}}$. Given a number $n \in \mathbb{N}$, we collect $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)_{i\in[n]} \overset{\text{i.i.d}}{\sim} \mathbb{P}_{\mathsf{train}}$ samples and let $(\hat{f}_n, \hat{g}_n)$ denote (any) empirical risk minimizer of the samples: $(\hat{f}_n, \hat{g}_n) \in \arg\min_{(f,g)\in\mathcal{F}\times\mathcal{G}} \hat{\mathcal{L}}_n(f, g)$, where

$$\hat{\mathcal{L}}_n(f, g) := \tfrac{1}{n}\textstyle\sum_{i=1}^n (f(\mathbf{x}_i) + g(\mathbf{y}_i) - \mathbf{z}_i)^2. \quad (2.2)$$

Distribution Shift. Although we have samples from $\mathbb{P}_{\mathsf{train}}$, we are primarily interested in the excess square loss under $\mathbb{P}_{\mathsf{test}}$. For simplicity, the body of this paper focuses on when the density ratios between these distributions are upper bounded; as discussed in Section 3.3, these conditions can be weakened considerably.

Definition 2.1. Define the *density ratio coefficients* $\nu_{x,y}, \nu_y \geq 1$ to be the smallest scalars such that for all measurable sets $A \subset \mathcal{X} \times \mathcal{Y}$ and $B \subset \mathcal{Y}$,

$$\mathbb{P}_{\mathsf{test}}[(\mathbf{x}, \mathbf{y}) \in A] \leq \nu_{x,y}\, \mathbb{P}_{\mathsf{train}}[(\mathbf{x}, \mathbf{y}) \in A]$$
$$\mathbb{P}_{\mathsf{test}}[\mathbf{y} \in B] \leq \nu_y\, \mathbb{P}_{\mathsf{train}}[\mathbf{y} \in B].$$

The interesting regime is where $\nu_{x,y}, \nu_y$ are finite. A standard covariate shift argument upper bounds the excess risk on $\mathbb{P}_{\mathsf{test}}$ by the joint density ratio, $\nu_{x,y}$, times the excess risk on $\mathbb{P}_{\mathsf{train}}$. Our aim is to show that much better bounds are possible. Specifically, if the class $\mathcal{F}$ is "smaller" than the class $\mathcal{G}$, then the excess risk on $\mathbb{P}_{\mathsf{test}}$ is *less sensitive* to shifts in the joint distribution (i.e., $\nu_{x,y}$) than it is to shifts in the $\mathbf{y}$-marginal (i.e., ν_y). Such an improvement is most interesting in the regime where $\nu_{x,y} \gg \nu_y$, which requires that $\mathbf{x}$ is not a measurable function of $\mathbf{y}$.

Controlling distribution shift via bounded density ratios is popular in the offline reinforcement learning, where such terms are called *concentrability coefficients* (Xie and Jiang, 2020; Xie et al., 2022). We stress that the uniform density ratio bounds in this section are merely for convenience; we discuss generalizations at length in Section 3.3.

Conditional Completeness. Notice that $(f_\star, g_\star)$ may not be identifiable in the model Eq. (2.1). The most glaring counterexample occurs when $\mathbf{x} = \mathbf{y}$, and $f_\star + g_\star \in \mathcal{F} \cap \mathcal{G}$. Then, $(f, g) = (f_\star + g_\star, \mathbf{0})$ and $(f, g) = (\mathbf{0}, f_\star + g_\star)$ are both optimal pairs of predictors. However, this setting is uninteresting for our purposes, since $\mathbf{x} = \mathbf{y}$ implies that $\nu_{x,y} = \nu_y$. On the other hand, when $\mathbf{x}$ and $\mathbf{y}$ are independent, the model is identifiable up to a constant offset, i.e., $(f_\star + c, g_\star - c)$ is an optimal pair. This line of reasoning suggests that the indentifiable part of $f_\star$ in Eq. (2.1) corresponds to the part of $\mathbf{x}$ that is orthogonal to $\mathbf{y}$. To capture this effect, we introduce the conditional *bias* of f given $\mathbf{y}$ under the *training distribution*:

$$\beta_f(\cdot) = \mathbb{E}_{\mathsf{train}}[(f - f_\star)(\mathbf{x}) \mid \mathbf{y} = \cdot]. \quad (2.3)$$

Note that this is a function of y, not x. One can check that $\mathcal{R}_{\mathrm{train}}(f, g) = 0$ if and only if $(f(\mathbf{x}), g(\mathbf{y})) = (f_\star(\mathbf{x}) + \beta_f(\mathbf{y}), g_\star(\mathbf{y}) - \beta_f(\mathbf{y}))$ with probability one over $(\mathbf{x}, \mathbf{y}) \sim \mathbb{P}_{\mathsf{train}}$. Note, in particular, that this requires β_f is almost surely (under $\mathbb{P}_{\mathsf{train}}$) equal to a measurable function of $\mathbf{x}$. This allows, for example, $(f, g) = (f_\star - c, g_\star + c)$ for constants $c \in \mathbb{R}$, and, in particular, $(f_\star, g_\star)$ meet these requirements since $\beta_{f_\star} = 0$.

We now introduce our final, and arguably only non-standard, assumption.

Assumption 2.3 (γ-Conditional Completeness)**.** There exists some $\gamma > 0$ such that, for any $(f, g) \in \mathcal{F}\times\mathcal{G}$ satisfying $\mathcal{R}_{\mathrm{train}}(f, g) \leq \gamma^2$, it holds that $g - \beta_f \in \mathcal{G}$.

Conditional completeness is somewhat non-intuitive but it is satisfied in some natural cases. We list them here informally, and defer formal exposition to Appendix A.1. First, as aluded to above, when $\mathbf{x} \perp \mathbf{y}$, $\beta_f(\mathbf{y})$ is constant in $\mathbf{y}$ and so conditional completeness holds as long as $\mathcal{G}$ is closed under affine translation. Second, it holds when $\mathcal{F}$ and $\mathcal{G}$ are linear classes and $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian; this follows since the conditional distribution $\mathbb{E}_{\mathsf{train}}[\mathbf{x} \mid \mathbf{y} = y]$ is linear in y. The latter example extends to nonparametric settings: conditional completeness holds if the conditional expectations $\mathbf{x} \mid \mathbf{y}$ are smooth and $\mathcal{G}$ contains correspondingly smooth functions.

The restriction to $\mathcal{R}_{\mathrm{train}}(f, g) \leq \gamma^2$ allows us to make the assumption compatible with the following, standard boundedness assumption (for otherwise we would need to have $g - k\beta_f \in \mathcal{G}$ for all $k \in \mathbb{N}$, see Remark A.1.)

Assumption 2.4 (Boundedness)**.** We assume that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$, $|f(\mathbf{x})|$ and $|g(\mathbf{y})|$ are uniformly bounded by some $B > 0$. For simplicity, we also assume $\mathcal{F}$ and $\mathcal{G}$ contain the zero predictor.

Notation. We use $a \lesssim b$ to denote inequality up to universal constants, and use $\mathcal{O}(\cdot)$ and $\widetilde{\mathcal{O}}(\cdot)$ as informal notation suppressing problem-dependent constants and logarithmic factors, respectively. A scalar-valued random variable is standard normal if $Z \sim \mathcal{N}(0, 1)$ and Rademacher

if Z is uniform on $\{-1, 1\}$. For $v = (v_1, \dots, v_n) \in \mathbb{R}^n$ and $q \in [1, \infty)$, define the normalized q-norms $\|v\|_{q,n} = (\frac{1}{n}\sum_{i=1}^n |v_i|^q)^{1/q}$ and $\|v\|_{\infty,n} = \|v\|_\infty = \max_{i\in[n]} |v_i|$. We let $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$ with elements $w \in \mathcal{W}$, so we can view classes $f \in \mathcal{F}, g \in \mathcal{G}$, and β_f as mappings with type $\mathcal{W} \to \mathbb{R}$. Given $h \in \mathcal{H}$ and a sequence $w_{1:n} \in \mathcal{W}^n$, define the evaluation vector $h[w_{1:n}] := (h(w_1), \dots, h(w_n)) \in \mathbb{R}^n$ and evaluated class $\mathcal{H}[w_{1:n}] := \{h[w_{1:n}] : h \in \mathcal{H}\} \subset \mathbb{R}^n$.

3. Results

All of our results follow from the same schematic: we argue that if $\mathcal{F}$ is simpler than $\mathcal{G}$, it is much easier to recover $f_\star$ than it is to recover $g_\star$, subject to the identifiability issues introduced by β_f. To express this, we introduce the per-function risks, for $\mathsf{e} \in \{\mathsf{train}, \mathsf{test}\}$:

$$\mathcal{R}_\mathsf{e}[f] := \mathbb{E}_\mathsf{e}[(f - f_\star - \beta_f)^2].$$
$$\mathcal{R}_\mathsf{e}[g; f] := \mathbb{E}_\mathsf{e}[(g - g_\star + \beta_f)^2].$$

Our schematic shows that $\mathcal{R}_{\mathrm{train}}[\hat{f}_n] \ll \mathcal{R}_{\mathrm{train}}[\hat{g}_n; \hat{f}_n]$, with precise convergence rates. The expression $\mathcal{R}_\mathsf{e}[f]$ reflects that f is identifiable only up to a bias, while $\mathcal{R}_\mathsf{e}[g; f]$ can be thought of as the residual error after accounting for the bias in f. A straightforward consequence of these definitions is the following risk decomposition:

Lemma 3.1. *Let $(f, g) \in \mathcal{F} \times \mathcal{G}$. Then, under Assms. 2.1 and 2.2, $\mathcal{R}_{\mathrm{train}}(f, g) = \mathcal{R}_{\mathrm{train}}[f] + \mathcal{R}_{\mathrm{train}}[g; f]$. Morever, $\mathcal{R}_{\mathsf{test}}(f, g) \le 2(\mathcal{R}_{\mathsf{test}}[f] + \mathcal{R}_{\mathsf{test}}[g; f])$. Therefore,*

$$\mathcal{R}_{\mathsf{test}}(f, g) \le 2(\nu_{x,y}\mathcal{R}_{\mathrm{train}}[f] + \nu_y \mathcal{R}_{\mathrm{train}}(f, g)). \quad (3.1)$$

Eq. (3.1) is the starting point for our results. By comparison, the standard distribution shift bound is

$$\mathcal{R}_{\mathsf{test}}(f, g) \le \nu_{x,y}\mathcal{R}_{\mathrm{train}}(f, g). \quad (3.2)$$

Hence, Eq. (3.1) leads to sharper estimates for ERM in the regime where $\nu_y \ll \nu_{x,y}$ and $\mathcal{R}_{\mathrm{train}}[\hat{f}_n] \ll \mathcal{R}_{\mathsf{test}}[\hat{f}_n, \hat{g}_n)$, i.e., when the shift in $\mathbf{y}$ is less than the shift in the joint distribution and when the estimate of $f_\star$ is more accurate than the estimate of $f_\star + g_\star$. The bulk of the analysis involves obtaining sharp bounds on $\mathcal{R}_{\mathrm{train}}[\hat{f}_n]$, this is sketched in Section 4. In the remainder of this section, we describe implications for various settings of interest.

3.1. Nonparametric Rates

We begin by demonstrating improvements in the *nonparametric regime*, where we measure the complexity of function classes by their metric entropies. Recall that an *ϵ-cover* of a set $\mathbb{V}$ in a norm $\|\cdot\|$ is a set $\mathbb{V}' \subset \mathbb{V}$ such that, for any $v \in \mathbb{V}$, there exists $v' \in \mathbb{V}'$ for which $\|v - v'\| \le \epsilon$. The *covering number* of $\mathbb{V}$ at scale ϵ in norm $\|\cdot\|$ is the minimal cardinality of an ϵ-cover, denoted $\mathscr{N}(\mathbb{V}, \|\cdot\|, \varepsilon)$. Metric entropies of function classes are defined via the logarithm of the covering number.

Definition 3.1 (Metric Entropy)**.** We define the q-norm metric entropy of a function class $\mathcal{H} : \mathcal{W} \to \mathbb{R}$ as $\mathcal{M}_q(\epsilon, \mathcal{H}) := \sup_n \sup_{w_{1:n}} \log \mathscr{N}(\mathcal{H}[w_{1:n}], \|\cdot\|_{q,n}, \varepsilon)$.

As in classical results in statistical learning theory, rates of convergence depend on function class complexity primarily through the *growth rate* of the metric entropy, i.e., how $\mathcal{M}_q(\epsilon, \mathcal{H})$ scales as a function of ϵ. We state our first main result informally, in line with this tradition.

Theorem 1 (Informal)**.** *Under Assm. 2.1-Assm. 2.4, with high probability, $\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n)$ is at most*

$$\widetilde{\mathcal{O}}\left(\nu_{x,y}(\mathrm{rate}_{n,2}(\mathcal{F}) + \mathrm{rate}_{n,\star}(\mathcal{G})^2) + \nu_y \mathrm{rate}_{n,2}(\mathcal{G})\right)$$

with high probability, where we define

$$\mathrm{rate}_{n,q}(\mathcal{H}) = \begin{cases} \frac{d}{n} & \mathcal{M}_q(\epsilon, \mathcal{H}) = \mathcal{O}\left(d \log(1/\epsilon)\right) \\ n^{-\frac{2}{2+p}} & \mathcal{M}_q(\epsilon, \mathcal{H}) = \mathcal{O}\left(\epsilon^{-p}\right),\ p \le 2 \\ n^{-\frac{1}{p}} & \mathcal{M}_q(\epsilon, \mathcal{H}) = \mathcal{O}\left(\epsilon^{-p}\right),\ p > 2 \end{cases}$$

and $\mathrm{rate}_{n,\star}(\mathcal{H}) = n^{-(1/2 \wedge 1/p)}$ for $\mathcal{M}_\infty(\epsilon, \mathcal{H}) = \mathcal{O}\left(\epsilon^{-p}\right)$.

A formal statement is given in Appendix E. As a preliminary point of comparison, the classical analysis would yield a bound of the form

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \ \le\ \widetilde{\mathcal{O}}\left(\nu_{x,y}(\mathrm{rate}_{n,2}(\mathcal{F}) + \mathrm{rate}_{n,2}(\mathcal{G}))\right),$$

which can be worse than the above bound when $\nu_y \ll \nu_{x,y}$ and $\mathrm{rate}_{n,\star}(\mathcal{G})^2 \ll \mathrm{rate}_{n,2}(\mathcal{G})$. It is crucial to note that in our bound, $\mathcal{G}$ and $\nu_{x,y}$ interact via the *squared* convergence rate for $\mathcal{G}$, which we expect to be significantly smaller than $\mathrm{rate}_{n,2}(\mathcal{G})$. This benign dependence on $\mathcal{G}$ is similar in spirit to results in orthogonal machine learning (c.f., Foster and Syrgkanis, 2019), where it arises from a similar phenomenon: the ability to recover $f_\star$ is only weakly impacted by the presence of $g_\star$. We discuss orthogonal ML in more detail in Section 3.4.

We should also mention one caveat to the bound: the ideal bound would replace $\mathrm{rate}_{n,\star}(\mathcal{G})^2$ with $\mathrm{rate}_{n,2}(\mathcal{G})^2$ which is smaller most notably when $\mathcal{G}$ is small. Unfortunately, our bound departs from this ideal in that $\mathrm{rate}_{n,\star}(\mathcal{G})$ involves the ℓ_∞ metric entropy and also that it does not exploit localization to go beyond the $n^{-1/2}$ rate for small classes. It is not clear if this discrepancy reflects a limitation in our analysis or is a fundamental limitation of ERM. On the other hand, localization can only improve the constant factors but not the dependence on n, since the $\mathrm{rate}_{n,\star}(\mathcal{G})$ term is *squared* and since $\mathrm{rate}_{n,2}(\mathcal{F})$ is always at least $1/n$.

3.2. Finite Function Classes

When $\mathcal{F}$ and $\mathcal{G}$ are finite function classes with $\log|\mathcal{F}| \le d_1$ and $\log|\mathcal{G}| \le d_2$, an application of Theorem 1 gives

the rate of $\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim \nu_{x,y} \cdot \frac{d_1+d_2}{n}$, which is precisely what one obtains via naive change of measure arguments. Although direct application of Theorem 1 does not yield improvements—precisely because of the lack of localization as discussed above— we *can* improve upon this bound with an additional *hypercontractivity* assumption, often popular in the statistical learning literature (Mendelson, 2015). We defer formal definitions, a formal theorem statement, and proofs to Appendix D; the following informal theorem summarizes our findings.

Theorem 2 (Informal)**.** *Under certain hypercontractivity conditions detailed in Appendix D, it holds with high probability that*

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim \frac{1}{n}\left(\nu_{x,y} d_1 + \nu_y d_2 + \nu_{x,y} d_2 \cdot \phi_n(d_1, d_2)\right),$$

where $\phi_n(d_1, d_2) = (\frac{d_2}{n})^{c_1} + (\frac{d_1}{d_2})^{c_2}$*, for constants* $c_1, c_2 > 0$ *depending on the hypercontractivity exponents.*

When $d_2 \gg d_1$, the bound replaces the dimension term $d_2\nu_{x,y}$ with $d_2\phi(d_1, d_2)\nu_{x,y} + \nu_y d_2$, a strict improvement when $\nu_y \ll \nu_{x,y}$ and $\phi(d_1, d_2) \ll 1$. The above bound can be extended to function classes with "parametric" metric entropy (Remark D.1). In all cases, $\phi_n(d_1, d_2) \geq \frac{d_2}{n}$, which is still weaker than an idealized version of Theorem 1 where $\mathrm{rate}_{n,2}(\mathcal{G})^2$ replaces $\mathrm{rate}_{n,\star}(\mathcal{G})^2$.

3.3. Refined Measures of Distribution Shift

The decomposition in Lemma 3.1 and all subsequent guarantees can be refined considerably. First, we can replace uniform bounds on the density ratios (Definition 2.1) with the following function-dependent quantities:

$$\nu_1 := \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_{\mathsf{test}}[(f - f_\star - \beta_f)^2]}{\mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta_f)^2]} \tag{3.3}$$

$$\nu_2 := \sup_{f \in \mathcal{F}, g \in \mathcal{G}} \frac{\mathbb{E}_{\mathsf{test}}[(g - g_\star - \beta_f)^2]}{\mathbb{E}_{\mathsf{train}}[(g - g_\star - \beta_f)^2]}, \tag{3.4}$$

Corollary 3.1. *Immediately from Lemma 3.1, it holds that* $\mathcal{R}_{\mathsf{test}}(f, g) \leq 2(\nu_1 \mathcal{R}_{\mathrm{train}}[f] + \nu_2 \mathcal{R}_{\mathrm{train}}(f, g))$

Both Theorem 1 and Theorem 2 continue to hold using ν_1 and ν_2 instead of $\nu_{x,y}$ and ν_y. Note that $\nu_1 \leq \nu_{x,y}$ and $\nu_2 \leq \nu_y$ always, but they can be much smaller as demonstrated by the follow upper bounds on ν_1.

Lemma 3.2. *Suppose* $\mathbf{x} \perp \mathbf{y}$ *under* $\mathbb{P}_{\mathsf{train}}$*. Then* $\nu_1 \leq \nu_x := \sup_{A \subset \mathcal{X}} \mathbb{P}_{\mathsf{test}}[\mathbf{x} \in A] / \mathbb{P}_{\mathsf{train}}[\mathbf{x} \in A]$.

Lemma 3.3. *Assume (a)* $\mathcal{X}$ *is a Hilbert space, (b) the functions* $f \in \mathcal{F}$ *are linear in* $\mathbf{x}$ *and (c) there are constants* $\nu_{\mathrm{lin}} > 0$ *such that, with* $\beta_x(\mathbf{y}) := \mathbb{E}_{\mathsf{train}}[\mathbf{x} \mid \mathbf{y}]$, $\mathbb{E}_{\mathsf{test}}[(\mathbf{x} - \beta_x(\mathbf{y}))(\mathbf{x} - \beta_x(\mathbf{y}))^{\mathsf{H}}] \preceq \nu_{\mathrm{lin}} \cdot \mathbb{E}_{\mathsf{train}}[(\mathbf{x} - \beta_x(\mathbf{y}))(\mathbf{x} - \beta_x(\mathbf{y}))^{\mathsf{H}}]$*. Then,* $\nu_1 \leq \nu_{\mathrm{lin}}$.

Appendix A.2 proves both lemmas. Importantly, ν_{lin} can be finite even when $\nu_{x,y}$ is infinite, e.g. if the distribution over $\mathbf{x}$ is discrete under $\mathbb{P}_{\mathsf{train}}$, but continuous under $\mathbb{P}_{\mathrm{test}}$.

Beyond Uniform Ratios. Eqs. (3.3) and (3.4) can be generalized further to allow for additive error.

Corollary 3.2. *Suppose that, for all* $f \in \mathcal{F}$ *and* $g \in \mathcal{G}$*,*

$$\mathbb{E}_{\mathsf{test}}[(f - f_\star - \beta_f)^2] \leq \nu_1 \mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta_f)^2] + \Delta_1$$
$$\mathbb{E}_{\mathsf{test}}[(g - g_\star - \beta_f)^2] \leq \nu_2 \mathbb{E}_{\mathsf{train}}[(g - g_\star - \beta_f)^2] + \Delta_2,$$

Then, immediately from Lemma 3.1,

$$\mathcal{R}_{\mathsf{test}}(f, g) \leq 2(\nu_1 \mathcal{R}_{\mathrm{train}}[f] + \nu_2 \mathcal{R}_{\mathrm{train}}(f, g) + \Delta_1 + \Delta_2).$$

This deceptively simple modification allows for situations when the density ratios between the test and train distributions are not uniformly bounded, or possibly even infinite. Appendix A.3 details the many consequences of this observation. We highlight a key one here:

Lemma 3.4. *Suppose Assm. 2.4 holds. Then, for* $\mathcal{R}_{\mathrm{train}}[f], \mathcal{R}_{\mathrm{train}}(f, g)$ *sufficiently small,*

$$\mathcal{R}_{\mathsf{test}}(f, g) \leq 8B\sqrt{\mathcal{R}_{\mathrm{train}}[f] \cdot \chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{x}, \mathbf{y}), \mathbb{P}_{\mathsf{train}}(\mathbf{x}, \mathbf{y}))} + 8B\sqrt{\mathcal{R}_{\mathrm{train}}(f, g) \cdot \chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{y}), \mathbb{P}_{\mathsf{train}}(\mathbf{y}))},$$

where $\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{x}, \mathbf{y}), \mathbb{P}_{\mathsf{train}}(\mathbf{x}, \mathbf{y}))$ *denotes the* χ^2 *divergence (see e.g. Polyanskiy and Wu (2022, Chapter 2)) between the joint distribution of* $(\mathbf{x}, \mathbf{y})$ *under test and train distributions, and* $\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{y}), \mathbb{P}_{\mathsf{train}}(\mathbf{y}))$ *denotes* χ^2 *divergence restricted to the marginals of* $\mathbf{y}$.

The above lemma is qualitatively similar to Corollary 3.1 and Lemma 3.1: If $\mathcal{R}_{\mathrm{train}}[f] \ll \mathcal{R}_{\mathrm{train}}(f, g)$ (as ensured by our analysis, under appropriate assumptions), then we ensure more resilience to the χ^2 divergence between the joint distributions of $(\mathbf{x}, \mathbf{y})$ than would naively be expected.

3.4. Comparison with Orthogonal ML

The style of our results is similar to those appearing in the literature on Neyman orthogonalization (also referred to as Double/Debiased ML or orthogonal statistical learning) (c.f., Chernozhukov et al., 2017; Mackey et al., 2018; Foster and Syrgkanis, 2019). At a high level, orthogonal ML considers a situation with an unknown pair $(f_\star, g_\star)$, where we are primarily interested in learning $f_\star$, referring to $g_\star$ as a nuisance function. The main difference in setting is that orthogonal ML leverages an auxiliary supervision mechanism to learn $g_\star$, which helps resolve confounding effects. Qualitatively, the rates are similar: as with our bound, orthogonal ML allows one to learn $f_\star$ with a quadratic dependence on the complexity $\mathcal{G}$. However quantatively the orthogonal ML bound is typically better in that (a) the complexity of $\mathcal{G}$ is measured via the ideal

$\mathrm{rate}_{n,2}(\mathcal{G})^2$ compared with our $\mathrm{rate}_{n,\star}(\mathcal{G})^2$, and (b) the distribution shift coefficient is ν_x compared with our $\nu_{x,y}$.

On the other hand, our bound is somewhat more practical, as it applies to ERM directly and does not require algorithmic modifications or an auxiliary supervision signal. The key difference here is that whereas orthogonal ML aims for *inference* – consistent recovery of $f_\star$ – we care only about the *prediction error* of $f_\star + g_\star$. Thus, we need not address the identifiability challenges present in orthogonal ML. As a consequence, we bypass algorithmic modifications that typically require more precise modeling of the data generating process, and which typically render orthogonal ML more susceptible to misspecification issues. Finally, we should note that in canonical settings for orthogonal learning, we can show that our main assumption, conditional completeness, holds. In this sense, our work shows that, in typically settings for orthogonal learning, one can obtain similar statistical improvements *with ERM alone* and *without auxiliary supervision*.

Please see Appendix A for a detailed discussion.

4. Analysis Overview

We begin this section with a formal precursor to Theorem 1 in terms of Dudley integrals (Dudley, 1967), stated as Theorem 3. The rest of the section provides an overview of the proof. Section 4.1 contains the necessary preliminaries, notably Rademacher and Gaussian complexities and their associated critical radii. Section 4.2 provides the roadmap for the proof of Theorem 3, focusing on our novel excess risk bound for $\mathcal{R}_{\mathrm{train}}[\hat{f}_n]$ in terms of a "cross critical radius" term. We bound this term in Section 4.3 via a Hölder style inequality for Rademacher complexity.

For convenience define the *centered classes* $\mathcal{F}_{\mathtt{cnt}} := \{f - \beta_f - f_\star : f \in \mathcal{F}\}$, $\mathcal{G}_{\mathtt{cnt}} := \{g - g_\star + \beta_f : f \in \mathcal{F}, g \in \mathcal{G}\}$ and $\mathcal{H}_{\mathtt{cnt}} := \{f + g - (f_\star + g_\star) : f \in \mathcal{F}, g \in \mathcal{G}\}$.

Formal Main Result. We define the Dudley functional, a standard measure of statistical complexity.

Definition 4.1 (Dudley Functional)**.** Let $\mathsf{rad}_q(\mathbb{V}) := \sup_{v\in\mathbb{V}} \|v\|_{q,n}$ be the q-norm radius and $\mathscr{M}_q(\mathbb{V};\cdot)$ be the metric entropy in the induced ℓ_q norm (Definition E.1). Given $\mathbb{V} \subset \mathbb{R}^n$ define *Dudley's chaining functional* (in the q-norm) as

$$\mathscr{D}_{n,q}(\mathbb{V}) := \inf_{\delta \le \mathsf{rad}_q(\mathbb{V})} \left(2\delta + \tfrac{4}{\sqrt{n}} \int_\delta^{\mathsf{rad}_q(\mathbb{V})} \sqrt{\mathscr{M}_q(\mathbb{V};\varepsilon/2)}\mathrm{d}\varepsilon\right).$$

Furthermore, given a function class $\mathcal{H}$ and letting $\mathcal{H}[r, w_{1:n}]$ denotes the empirically localized class (Definition 4.2 below), define *the Dudley critical radius*

$$\delta_{n,\mathscr{D}}(\mathcal{H}, c) := \inf\left\{r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le \tfrac{r^2}{2c}\right\},$$

We now state the formal version of our main result. We obtain Theorem 1 by bounding the Dudley functionals using standard statistical learning arguments.

Theorem 3. *Suppose Assms. 2.1 to 2.3 hold. Let* $\sigma_B := \max\{B, \sigma\}$*, let* ν_1, ν_2 *be as in Eqs. (3.3) and (3.4), and let* c_1 *be a sufficiently small universal constant. Then if* $\delta_{n,\mathscr{D}}(\mathcal{H}_{\mathtt{cnt}}, \sigma_B)^2 + \frac{\sigma_B^2 \log(1/\delta)}{n} \le c_1\gamma$*, it holds that probability at least* $1-\delta$*,* $\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n)$ *is at most*

$$\lesssim \nu_1 \left(\delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, \sigma_B)^2 + \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{G}_{\mathtt{cnt}}[w_{1:n}])^2\right) + \nu_2 \cdot \delta_{n,\mathscr{D}}(\mathcal{H}_{\mathtt{cnt}}, \sigma_B)^2 + \frac{(\nu_1+\nu_2)\sigma_B^2 \log(1/\delta)}{n}.$$

4.1. Learning-Theoretic Preliminaries

We state all definitions for a general class of functions $\mathcal{H}$ mapping $\mathcal{W} \to \mathbb{R}$. We define two key notions of *localized* and *product classes*.

Definition 4.2 (Product and Localized Classes)**.** Let $\mathcal{H}, \mathcal{H}' : \mathcal{W} \to \mathbb{R}$. Define the *(Hadamard) product class* $\mathcal{H} \odot \mathcal{H}' := \{h \cdot h' : h \in \mathcal{H}, h \in \mathcal{H}'\}$. Given sequence $w_{1:n} \in \mathcal{W}^n$, define the *empirically localized* class $\mathcal{H}[r, w_{1:n}] := \{h \in \mathcal{H} : \frac{1}{n}\sum_{i=1}^n h(w_i)^2 \le r\}$ and *population localized* class $\mathcal{H}(r) := \{h \in \mathcal{H} : \mathbb{E}_{\mathsf{train}}[h^2] \le r\}$.

Next, we define the standard Rademacher and Gaussian complexities and associated quantities (c.f., Rakhlin, 2022; Wainwright, 2019; Bartlett et al., 2005). For convenience, we state these quantities for a set of n-length vectors $\mathbb{V} \subset \mathbb{R}^n$ and then instantiate the definition to obtain function class variants.

Definition 4.3 (Rademacher and Gaussian Complexities: Sets)**.** Let $n \in \mathbb{N}$, and let $\boldsymbol{\varepsilon}_{1:n}$ and $\boldsymbol{\xi}_n$ denote i.i.d. sequences of Rademacher and standard Normal random variables, respectively. The Rademacher and Gaussian complexities of a subset $\mathbb{V} \subset \mathbb{R}^n$ are defined as $\mathscr{R}_n(\mathbb{V}) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{v\in\mathbb{V}} \sum_{i=1}^n \boldsymbol{\varepsilon}_i v_i$ and $\mathscr{G}_n(\mathbb{V}) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\xi}} \sup_{v\in\mathbb{V}} \sum_{i=1}^n \boldsymbol{\xi}_i v_i$.

Gaussian and Rademacher complexities of function classes can be defined in terms of Definition 4.3. For example, we may consider $\mathscr{R}_n(\mathcal{H}[w_{1:n}])$, or localized variants like $\mathscr{R}_n(\mathcal{H}[r, w_{1:n}])$. For the latter, we define the *critical radius* quantities, which are central to localization arguments in statistical learning (Bartlett et al., 2005).

Definition 4.4 (Critical Radii)**.** We define the following worst-case *critical radii*:
$\delta_{n,\mathscr{R}}(\mathcal{H}, c) := \inf\left\{r : \sup_{w_{1:n}} \mathscr{R}_n(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2c}\right\}$,
$\delta_{n,\mathscr{G}}(\mathcal{H}, c) := \inf\left\{r : \sup_{w_{1:n}} \mathscr{G}_n(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2c}\right\}$.

The following lemma verifies that the Rademacher and Gaussian complexities are upper bounded by the Dudley functional (the proof is standard, but see also Appendix B.7 for completeness.)

Lemma 4.1. *For any $\mathbb{V} \subset \mathbb{R}^n$, we have $\mathscr{G}_n(\mathbb{V}) \vee \mathscr{R}_n(\mathbb{V}) \le \mathscr{D}_{n,2}(\mathbb{V})$, and hence, for all $c > 0$, $\delta_{n,\mathscr{R}}(\mathcal{H}, c) \vee \delta_{n,\mathscr{G}}(\mathcal{H}, c) \le \delta_{n,\mathscr{D}}(\mathcal{H}, c)$.*

4.2. Proof Overview of Theorem 3.

We begin with the following generic upper bound on the joint risk of $\mathcal{R}_{\mathrm{train}}[\hat{f}_n, \hat{g}_n]$, proved in Appendix B.2.

Proposition 4.2. *With probability at least $1-\delta$, we have that $\mathcal{R}_{\mathrm{train}}[\hat{g}_n; \hat{f}_n] \le \mathcal{R}_{\mathrm{train}}(\hat{f}_n, \hat{g}_n) \lesssim \gamma_n(\delta)^2$, where we define $\gamma_n(\delta)^2 := \delta_{n,\mathscr{R}}(\mathcal{H}_{\mathtt{cnt}}, B)^2 + \delta_{n,\mathscr{G}}(\mathcal{H}_{\mathtt{cnt}}, \sigma)^2 + \frac{(B^2+\sigma^2)\log(1/\delta)}{n}$. Hence, $\hat{g}_n \in \mathcal{G}_{\mathtt{cnt}}(\gamma_n(\delta))$.*

The localized Gaussian complexity of the class $\mathcal{H}_{\mathtt{cnt}}$, $\delta_{n,\mathscr{G}}(\mathcal{H}_{\mathtt{cnt}}, \sigma)^2$, appears in the sharpest analyses of ERM. The dependence on $\delta_{n,\mathscr{R}}(\mathcal{H}_{\mathtt{cnt}}, B)^2$ is suboptimal in general (see, e.g. Rakhlin (2022, Chapter 21)), but is convenient and essentially sharp in the regime where $\sigma^2 \gtrsim B^2$. However, to take advantage of Eq. (3.1), we require sharper control over $\mathcal{R}_{\mathrm{train}}[\hat{f}_n]$. This involves a novel term, unique to our setting; the *cross-critical radius*.

Definition 4.5 (Cross Critical Radii)**.** Given the classes $\mathcal{F}_{\mathtt{cnt}}$, and another class $\mathcal{H}$, we define $\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{H}) := \inf\left\{ r : \sup_{w_{1:n}} \mathscr{R}_n(\mathcal{F}_{\mathtt{cnt}}[r, w_{1:n}] \odot \mathcal{H}[w_{1:n}]) \le \frac{r^2}{2} \right\}$.

The cross-critical radius measures the complexity of products $\tilde{f} \cdot h$, where $\tilde{f} \in \mathcal{F}_{\mathtt{cnt}}$ and $h \in \mathcal{H}$, and thus captures the extent to which h can obfuscate recovery of $f_\star$. We invoke the cross-critical radii with $\mathcal{H} = \mathcal{G}_{\mathtt{cnt}}(\gamma)$, for some γ.

It is crucial that the localization $r > 0$ is *only* on the class $\mathcal{F}_{\mathtt{cnt}}$ and not on the class $\mathcal{H}$. With the cross-critical radius in hand, the following is proved in Appendix B.3.

Proposition 4.3. *Suppose that $(\mathcal{F}, \mathcal{G})$ satisfy γ-conditional completeness. Then, whenever $\mathcal{R}_{\mathsf{train}}[\hat{g}_n; \hat{f}_n] \le \gamma$, the following holds with probability at least $1-\delta$,*

$$\mathcal{R}_{\mathsf{train}}[\hat{f}_n] \lesssim \delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma))^2 + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta_{n,\mathscr{G}}(\mathcal{F}_{\mathtt{cnt}}, \sigma)^2 + \frac{(\sigma^2+B^2)\log(1/\delta)}{n}.$$

Upper bounding Rademacher and Gaussian critical radii by the Dudley radius, and using $\sigma_B = \max\{\sigma, B\}$, we have that with probability $1-\delta/2$, $\mathcal{R}_{\mathrm{train}}[\hat{g}_n; \hat{f}_n] \lesssim \delta_{n,\mathscr{D}}(\sigma_B, \mathcal{H}_{\mathtt{cnt}})^2 + \sigma_B^2 \log(1/\delta)/n$. Hence, if $\delta_{n,\mathscr{D}}(\sigma_B, \mathcal{H}_{\mathtt{cnt}})^2 + \sigma_B^2 \log(1/\delta)/n \le c_1\gamma$ for a small enough $c_1 > 0$, Proposition 4.3 yields that with probability $1-\delta/2$, $\mathcal{R}_{\mathsf{train}}[\hat{f}_n]$ is at most

$$\lesssim \delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma))^2 + \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, \sigma_B)^2 + \frac{\sigma_B^2 \log\frac{1}{\delta}}{n} \le \delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}})^2 + \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, \sigma_B)^2 + \frac{\sigma_B^2 \log\frac{1}{\delta}}{n}.$$

Invoking Corollary 3.1 and summing these bounds, with probability at least $1-\delta$, $\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n)$ is at most

$$\lesssim \nu_1(\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}})^2 + \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, \sigma_B)^2) + \nu_2 \delta_{n,\mathscr{D}}(\sigma_B, \mathcal{H}_{\mathtt{cnt}})^2 + \frac{(\nu_1+\nu_2)\sigma_B^2 \log(1/\delta)}{n}, \quad (4.1)$$

The key step is to now apply Corollary 4.1, stated in the following section, to upper bound the cross critical radius: $\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}})^2 \le \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, 2B)^2 + 16 \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{G}_{\mathtt{cnt}}[w_{1:n}])^2$. By Lemma C.4 and the bound $B \le \sigma_B$, $\delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, 2B)^2 \lesssim \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 \le \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, \sigma_B)^2$. Combining with Eq. (4.1) concludes the proof of Theorem 3. □

4.3. Controlling Cross Critical Radius via a Hölder-Inequality for Dudley's integral

Recall Lemma 4.1, which restates the well-known fact that the Rademacher and Gaussian complexities of a function class can be upper bounded by Dudley functional defined in Definition 4.1 (c.f., Dudley, 1967; Wainwright, 2019, Chapter 5). We establish a Hölder style generalization of this upper bound. In what follows, given $p, q \in [2, \infty]$, we say (p, q) are *square Hölder conjugates* if $(p/2, q/2)$ are regular Hölder conjugates, i.e. $\frac{2}{p} + \frac{2}{q} = 1$. Examples include $(p, q) = (2, \infty)$, $(p, q) = (4, 4)$, and $p, q = (\infty, 2)$. If $v, u \in \mathbb{R}^n$ are two vectors, then Hölder's inequality implies that for any square Hölder conjugates p, q, $\|v \odot u\|_{2,n} \le \|v\|_{2,p} \cdot \|u\|_{q,n}$. It may be tempting to generalize Lemma 4.1 to product classes via

$$\mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) \lesssim \mathscr{D}_{n,p}(\mathbb{V}) \cdot \mathscr{D}_{n,q}(\mathbb{U}), \quad (4.2)$$

where we recall $\mathbb{V} \odot \mathbb{U} := \{v \odot u, v \in \mathbb{V}, u \in \mathbb{U}\}$. Our key result is that Eq. (4.2) can be sharpened considerably. The following technical result is proved in Appendix B.6.

Proposition 4.4 (Dudley Estimate for Hadamard Products (Sets))**.** *Let $p, q \in [2, \infty]$ satisfy $1/p + 1/q \le 1/2$, and (for simplicitly) suppose $\mathbf{0} \in \mathbb{V} \cap \mathbb{W}$. Then,*

$$\mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) \le \mathsf{rad}_q(\mathbb{U}) \mathscr{D}_{n,p}(\mathbb{V}) + \mathsf{rad}_p(\mathbb{V}) \mathscr{D}_{n,q}(\mathbb{U}).$$

The same bound holds for $\mathscr{R}_n$ replaced by any process defined where the ε_i are 1-subGaussian variables (e.g. Gaussian complexity $\mathscr{G}_n$).

Notice that rather than having $\mathscr{D}_{n,p}(\mathbb{V})$ and $\mathscr{D}_{n,q}(\mathbb{U})$ multiply each other as in Eq. (4.2), each is only multiplied by the (Hölder square conjugate) radius term. This is in general considerably sharper, as typically $\mathsf{rad}_p(\mathbb{V}) \ll \mathscr{D}_{n,p}(\mathbb{V})$ unless $\mathbb{V}$ is exceedingly small. By taking $\mathbb{U} = \{(1, 1, \ldots, 1) \in \mathbb{R}^n\}$, Proposition 4.4 implies the standard Dudley bound, Lemma 4.1, as a corollary (see Appendix B.7). We now use the above proposition to upper bound the cross-critical radius.

Corollary 4.1 (Generic Cross-Critical Radius Bound)**.** *For any class $\mathcal{H}$, it holds that* $\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}};\mathcal{H})^2 \leq \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}},2B)^2 + 16\sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}])^2$.

Proof of Corollary 4.1. Recall the defintion of $\delta_{n,\mathrm{cross}}$ (Definition 4.5) By Proposition 4.4 with $(p,q)=(2,\infty)$,

$$\begin{aligned}
&\mathscr{R}_n(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]\odot\mathcal{H}[w_{1:n}]) \\
&\leq \mathsf{rad}_2(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}])\mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}]) \\
&\quad + \mathsf{rad}_\infty(\mathcal{H})\mathscr{D}_{n,2}(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]) \\
&\leq r\mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}]) + B\mathscr{D}_{n,2}(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]),
\end{aligned}$$

where we use that $\mathsf{rad}_2(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]) \leq r$ by the definition of localization. In particular, if r satisfies $\frac{r^2}{4} \geq B\sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}])$ and $\frac{r}{4} \geq \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}])$, then $\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}};\mathcal{H}) \leq r$. Thus, $\delta_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}};\mathcal{H})$ is at most the maximum of $\inf\{r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]) \leq \frac{r^2}{4B}\}$ and $4\sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}])$, which is at most the maximum of $\inf\{r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{F}_{\mathtt{cnt}}[r,w_{1:n}]) \leq \frac{r^2}{4B}\} = \delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}},2B)$ and $4\sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}])$. The bound follows by squaring. □

Remark 4.1. Because we consider the ∞-norm Dudley integral of $\mathcal{G}_{\mathtt{cnt}}$, it is hard to take advantage of localization of $\mathcal{G}_{\mathtt{cnt}}$ at $\mathcal{G}_{\mathtt{cnt}}(\gamma)$ in the $\mathcal{L}_2$ norm. The absence of localization leads to the suboptimal dependence on leading constants compared to what is obtained through Double ML (Foster and Syrgkanis, 2019). Appendix D shows that, for finite-function classes, one can take advantage of localization with strong hypercontractivity assumptions.

5. Experiments

In this section we present experiments to validate our theoretical findings and demonstrate how the conceptual takeaways—that predictive models are more resilient to distribution shifts in simple features—applies to a broad range of practical settings. All of our experiments have a similar form: we (a) identify simple and complex features and justify these choices and (b) measure how the performance of a predictive model changes with distribution shifts in these features. We experiment with neural network models on tasks ranging from synthetic regression problems to computer vision benchmarks to imitation learning in a robotics simulator. We take the following operational definition of simplicity:

> *Feature 1 is simpler than feature 2 if the generalization error, without distribution shift, on a predictive task involving feature 1 is smaller than that for an analogous task involving feature 2.*

We believe that this is the correct empirical correlate for the complexity measures adopted by our theoretical results. Across domains, we consistently find that *predictive models are more resilient to shifts in simpler features, thus defined.* We now summarize the experimental results, deferring details to Appendix F. In all experiments, we report average performance and standard error across 4 replicates.

Synthetic Regression with Additive Structure. To closely mirror our theory, we predict $\mathbf{z} = f_\star(\mathbf{x}) + g_\star(\mathbf{y})$ from an input $(\mathbf{x},\mathbf{y})$, where $f_\star : \mathbb{R}^{d_x} \to \mathbb{R}$ and $g_\star : \mathbb{R}^{d_y} \to \mathbb{R}$ are randomly initialized 2-layered multi-layer perceptions (MLP) having hidden dimensions d_f and d_g, respectively. We sample $\mathbf{x}$ and $\mathbf{y}$, respectively, independently from Gaussian mixtures $p_x = \mathcal{N}(\mathbf{1}_{d_x}, I) + (1-p_x)\mathcal{N}(-\mathbf{1}_{d_x}, \mathbf{I}_{d_x})$ and $p_y = \mathcal{N}(\mathbf{1}_{d_y}, \mathbf{I}_{d_y}) + (1-p_y)\mathcal{N}(-\mathbf{1}_{d_y}, \mathbf{I}_{d_y})$ where p_x, p_y are mixing probabilties, and $\mathbf{1}_d \in \mathbb{R}^d$ and $\mathbf{I} \in \mathbb{R}^{d\times d}$ are the all ones vector and identity matrix. To make $\mathbf{x}$ simpler, we either have $d_x < d_y$ or $d_f < d_g$: Appendix F.1 shows that the auxiliary task of predicting $f_\star(\mathbf{x})$ has lower generalization error than that of predicting $g_\star(\mathbf{y})$. We train a predictor $\hat{z} = f_\theta(x) + g_\theta(y)$ to minimize mean-square error (MSE) under a training distribution with $p_x = p_y = .01$. We then measure the MSE on shifted distributions, where we hold the mixing probability of one of $\{\mathbf{x},\mathbf{y}\}$ fixed, and vary the other in the range $\{0.1, 0.2, 0.5, 0.9, 0.99\}$. Corroborating our theory, MSE declines less with shift in p_x than with shift in p_y (Figure 1). Appendix F.1 shows similar results with predictor $\hat{z} = h_\theta(x,y)$ (using concatenated features $(\mathbf{x},\mathbf{y})$) and contains further implementation details.

Waterbird & Functional Map of World datasets. We next test our hypothesis on the paradigmatic Waterbird dataset (Sagawa et al., 2019). The predictive task is to classify images of birds as waterbirds or landbirds, against a background of either land or water. We consider `birdtype` (resp. `background`) as the complex (resp. simple) feature. This choice is both intuitive and consistent with our definition of simplicity: Appendix F.2 shows that, in the absense of distribution shift, the auxiliary task of predicting `background` has lower generalization error than that of predicting `birdtype`. We then test (Figure 1) the prediction accuracy of `birdtype` under shifts in the proportions of `background` and `birdtype`, finding prediction accuracy degrades less for the former than the latter. See Appendix F.2 for details. Appendix F.3 applies the same methodology to the Functional Map of World (FMoW) dataset (Koh et al., 2021), with similar findings.

Logical operators on CelebA dataset. The CelebA dataset (Liu et al., 2015) consists of celebrity faces labeled with the presence or absence of 40 different attributes (e.g., baldness, mustache). Here we re-purpose CelebA to learn logical operators OR and XOR for two at-

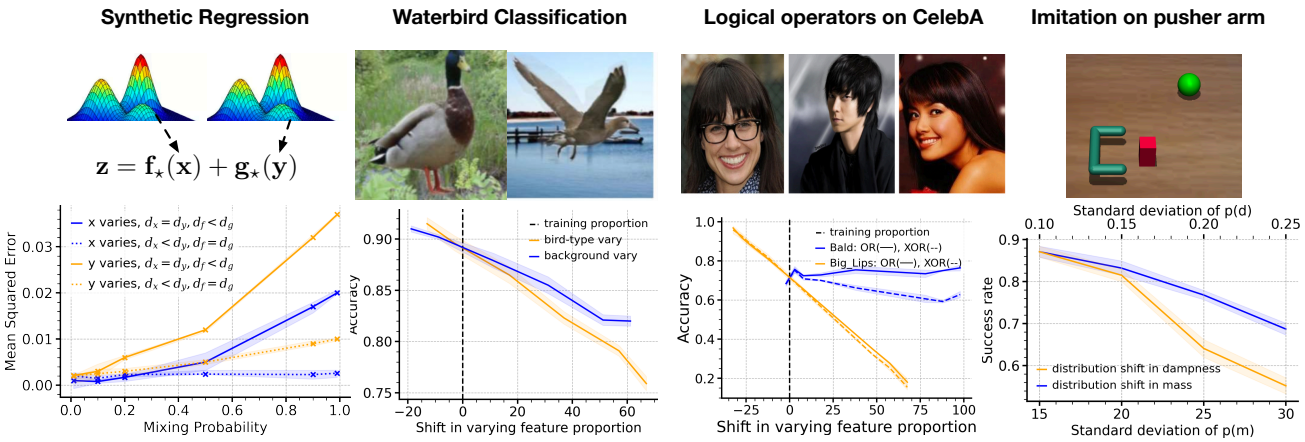

Figure 1. **Testing resiliency of learned predictive models**. We calculate the performance of learned models as we shift the distribution of either *complex* features (orange color) or *simple* features (blue color). To shift distribution of $\{\mathbf{x}, \mathbf{y}\}$ in synthetic regression, we vary the mixing probabilities of their distributions. To shift distribution of a feature in Waterbirds and CelebA datsets, we vary its proportion. To shift distribution of a feature in robotic pusher arm control, we increase standard deviation of its sampling distribution. Corroborating our theoretical expectations, we show that these models are more robust to distribution shift in *simple* features.

tributes. We first train and test a multi-head binary classifier that detects presence of 40 different attributes, with one head per attribute, on images from the CelebA "standard training set" (CelebA-STS). We select `bald` as "simple" due to its low generalization error and `big_lips` as "complex" due to its larger generalization error. In Figure 1, we predict targets $f_{\mathsf{OR}} =$ `bald` OR `big_lips` and $f_{\mathsf{XOR}} =$ `bald` XOR `big_lips`, training on CelebA-STS, but testing on distributions where the proportions of `bald` and `big_lips` are varied (Appendix F.4). Results show greater resilience to shift in the simpler feature `bald` than the complex feature `big_lips`. Further details are deferred to Appendix F.4, where we perform the same experiment for (`pale_skin`, `narrow_eyes`) with similar findings.

Imitation Learning for Pusher Control. In Pusher Control simulator, we learn an agent that controls a robotic arm to push an object to its goal location. We fix the goal and starting object locations across episodes. We vary object `mass` m ($m \sim \mathcal{N}(60, 15)$) and joint `dampness` d ($d \sim \mathcal{N}(0.5, 0.1)$) of the robot. The agent observes $\{m, d\}$ at beginning of every episode and aims to learn an optimal policy condition on these variables. To determine the simpler feature, we measure generalization error on the auxiliary task of predicting next-step dynamics where one of $\{m, d\}$ is held fixed and the other is drawn from a distribution that is fixed across training and testing. This methodology ascribes m as the simpler feature and d as complex (Appendix F.5). We train a policy $\pi_\theta(a|s, m, d)$ using 1000 expert trajectories where each trajectory contains a new m and d sampled from $\mathcal{N}(60, 15)$ and $\mathcal{N}(0.5, 0.1)$ respectively. We then shift distribution of m and d by increasing their standard deviation, one at a time while keeping the distribution of the other factor fixed, and test policy $\pi_\theta(a|s, m, d)$. We show the results in Figure 1 and observe that the success rate of $\pi_\theta(a|s, m, d)$ deteriorates less when we shift distribution of m while keeping distribution of d fixed. Thus, we show that the policy is more resilient to distribution shift in the simpler feature. Appendix F.5 contains further details, including the precise distributions used to determine m as the simpler feature.

6. Discussion

This paper sheds new light on the issue of spurious correlation that arises when considering out of distribution generalization. We discover that predictive models are more resilient to distribution shift in simpler features, which we capture via notions of statistical capacity in our experiments and via generalization when predicting the feature itself in our experiments. We find that, in most of our experiments, this latter operational notion is predictive of how deep learning models behave under heterogeneous distribution shift. We hope that our work inspires future efforts toward a fine-grained theoretical and experimental understanding of distribution shift in modern machine learning.

Acknowledgements

MS acknowledges support from Amazon.com Services LLC grant; PO# 2D-06310236. AA and PA acknowledge supported from a DARPA Machine Common Sense grant, a MURI grant from the Army Research Office under the Cooperative Agreement Number W911NF-21-1-0097, and an MIT-IBM grant. The authors thank Adam Block for his assistance in navigating the relevant learning theory literature.

References

Anurag Ajay, Abhishek Gupta, Dibya Ghosh, Sergey Levine, and Pulkit Agrawal. Distributionally adaptive meta reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893*, 2019.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 2005.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.

Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 2002.

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2002.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 2017.

Kefan Dong and Tengyu Ma. First steps toward understanding the extrapolation of nonlinear models to unseen domains. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=7wrq3vHcMM.

Richard M Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1967.

Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv:1901.09036*, 2019.

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta-reinforcement learning of structured exploration strategies. *Advances in Neural Information Processing Systems*, 2018.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv:1812.05905*, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*, 2017.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavnass, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pages 6164–6174. PMLR, 2021.

Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, 2015.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *arXiv:2205.02986*, 2022.

Lester Mackey, Vasilis Syrgkanis, and Ilias Zadik. Orthogonal machine learning: Power and limitations. In *International Conference on Machine Learning*, 2018.

Shahar Mendelson. Learning without concentration. *Journal of the ACM*, 2015.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, 2021.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, 2022.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, 2022.

Hamed Rahimian and Sanjay Mehrotra. Distributionally robust optimization: A review. *arxiv:1908.05659*, 2019.

Alexander Rakhlin. IDS.160 – Mathematical Statistics: A non-asymptotic approach, 2022. URL http://www.mit.edu/~rakhlin/courses/mathstat/rakhlin_mathstat_sp22.pdf.

Peter M Robinson. Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 1988.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shibani Santurkar, Dimitris Tsipras, and Aleksander Madry. Breeds: Benchmarks for subpopulation shift. *arXiv:2008.04859*, 2020.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 2018.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision*, 2016.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 2020.

Aad W van der Vaart and Jon A Wellner. *Weak convergence*. Springer, 1996.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press, 2018.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M Kakade. The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*, 2022.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A. Discussion of Assumptions

In this section we provide additional details about our main assumptions, conditional completeness. As this is closely related to orthogonal ML, we also provide some additional context and comparisons.

A.1. On Conditional Completeness

Our results hinge on the conditional completeness assumption, where $\mathcal{G}$ is expressive enough to capture the conditional bias functions β_f. To clarify when this assumption may hold, we discuss an example.

Partially Linear Regression. First, let us consider a paradigmatic model in econometrics, statistics, and causal inference. This model, known as the partial linear regression (PLR) model (Chernozhukov et al., 2017; Robinson, 1988), specifies a joint distribution over tuples $(\mathbf{z}, \mathbf{x}, \mathbf{y})$ via the structural equations:

$$\begin{aligned}\mathbf{z} &= \langle \mathbf{x}, f_\star \rangle + h_1(\mathbf{y}) + \varepsilon \\ \mathbf{x} &= h_0(\mathbf{y}) + \tau \\ \mathbf{y} &\sim P, \mathbb{E}[\varepsilon \mid \mathbf{x}, \mathbf{y}] = 0, \mathbb{E}[\tau \mid \mathbf{y}] = 0\end{aligned}$$

Here $\mathbf{x}$ belongs to a (finite dimensional) vector space, say $\mathbb{R}^d$, and $f_\star$ is a linear function, so we use $f_\star$ both to describe the mapping and the vector itself. In the learning setting for PLR, we are given access to function class $\mathcal{H}_0, \mathcal{H}_1$ such that $h_0 \in \mathcal{H}_0, h_1 \in \mathcal{H}_1$, and we assume that $f_\star$ has bounded norm, say $\|f_\star\| \leq B$.

This model is amenable to our techniques whenever $\mathcal{H}_1$ consists of linear projections of $\mathcal{H}_0$, i.e.,

$$\mathcal{H}_0 := \{\mathbf{y} \mapsto \langle v, h_1(\mathbf{y}) \rangle : v \in \mathbb{R}^d, h_1 \in \mathcal{H}_1\}.$$

Clearly we can apply ERM with $\mathcal{F}$ as the linear class and $\mathcal{G} = \mathcal{H}_0$. But we must verify that conditional completeness holds. This follows because, for any f, we have

$$\beta_f(\mathbf{y}) := \mathbb{E}[\langle f - f_\star, \mathbf{x} \rangle \mid \mathbf{y}] = \langle f - f_\star, \mathbb{E}[\mathbf{x} \mid \mathbf{y}] \rangle = \langle f - f_\star, h_0(\mathbf{y}) \rangle \in \mathcal{G}.$$

Thus, our results demonstrate favorable guarantees for estimating $f_\star$ via ERM in this setting.

Remark A.1 (Compatibility of conditional completeness and boundedness)**.** If conditional-completeness were stipulated as a global condition, i.e. for all $(f, g) \in \mathcal{F} \times \mathcal{G}$, $g - \beta_f \in \mathcal{G}$, then necessarily $(f, g - k\beta_f) \in \mathcal{F} \times \mathcal{G}$ for all $k \in \mathbb{N}$. Therefore, $\mathcal{G}$ would not in general consist only of functions which are uniformly bounded (except in the special case where $\beta_f = 0$ for all $f \in \mathcal{F}$). By imposing the restriction $\mathcal{R}_{\text{train}}(f, g) \leq \gamma^2$, we avoid this pathology because, unles $\beta_f \equiv 0$, $\lim_{k \to \infty} \mathcal{R}_{\text{train}}(f, g - k\beta_f) = \infty$.

A.2. Proof of Lemmas 3.2 and 3.3

Proof of Lemma 3.2. Suppose that if $\mathbf{x} \perp \mathbf{y}$ under $\mathbb{P}_{\mathsf{train}}$. Then $\beta_f(y) = \mathbb{E}[f(\mathbf{x}) - f_\star(\mathbf{x}) \mid \mathbf{y} = y] = \mathbb{E}[f(\mathbf{x}) - f_\star(\mathbf{x})]$ is a constant in y.

$$\begin{aligned}\nu_1 &:= \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_{\mathsf{test}}[(f - f_\star - \beta_f)^2]}{\mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta_f)^2]} \\ &\leq \sup_{f \in \mathcal{F}, c \in \mathbb{R}} \frac{\mathbb{E}_{\mathsf{test}}[(f - f_\star - c)^2]}{\mathbb{E}_{\mathsf{train}}[(f - f_\star - c)^2]} \\ &\leq \sup_{A \subset \mathcal{X}} \frac{\mathbb{P}_{\text{test}}[\mathbf{x} \in A]}{\mathbb{P}_{\mathsf{train}}[\mathbf{x} \in A]} = \nu_x,\end{aligned}$$

as the functions $f - f_\star - c$ are functions of $\mathbf{x}$ alone. □

Proof of Lemma 3.3. By linearity (assumption (a) of Lemma 3.3, we can write $f(\mathbf{x}) = \langle v, \mathbf{x} \rangle$ and $f_\star(\mathbf{x}) = \langle v^\star, \mathbf{x} \rangle$ for

some $v, v^\star \in \mathbb{H}$. Then, setting $\beta_x(\mathbf{y}) = \mathbb{E}_{\mathsf{train}}[\mathbf{x} \mid \mathbf{y}]$,

$$\begin{aligned}
&\mathbb{E}_{\mathsf{test}}[(f(\mathbf{x}) - f_\star(\mathbf{x}) - \beta_f(\mathbf{x}))^2] = \mathbb{E}_{\mathsf{test}}[\langle v - v^\star, \mathbf{x} - \beta_x(\mathbf{y})\rangle^2] \\
&\le \nu_{\mathrm{lin}} \mathbb{E}_{\mathsf{train}}[\langle v - v^\star, \mathbf{x} - \beta_x(\mathbf{y})\rangle^2] && \text{(Lemma 3.3, assumption (c))} \\
&= \nu_{\mathrm{lin}} \mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f_\star(\mathbf{x}) - \underbrace{\mathbb{E}_{\mathsf{train}}[f(\mathbf{x}) - f_\star(\mathbf{x}) \mid \mathbf{y}]}_{=\beta_f(\mathbf{y})})^2],
\end{aligned}$$

as needed. □

A.3. Beyond Uniform Density Bounds

In this section, we consider a generalization of Eqs. (3.3) and (3.4) to allow for additive slack. We show that this allows for generalizations of Lemma 3.1 and Corollary 3.1 which accomodate density ratios which are possibly unbouned, or even take the value ∞ with positive probability (Appendix A.3.1). Finally, we show that our guarantees imply guarantees when the χ^2 divergence (or more generally, power divergence) are bounded (Appendix A.3.2), thereby establishing the proof of Lemma 3.4. We begin by restating Corollary 3.2.

Corollary 3.2. *Suppose that, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,*

$$\begin{aligned}
\mathbb{E}_{\mathsf{test}}[(f - f_\star - \beta_f)^2] &\le \nu_1 \mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta_f)^2] + \Delta_1 \\
\mathbb{E}_{\mathsf{test}}[(g - g_\star - \beta_f)^2] &\le \nu_2 \mathbb{E}_{\mathsf{train}}[(g - g_\star - \beta_f)^2] + \Delta_2,
\end{aligned}$$

Then, immediately from Lemma 3.1,

$$\mathcal{R}_{\mathsf{test}}(f, g) \le 2(\nu_1 \mathcal{R}_{\mathrm{train}}[f] + \nu_2 \mathcal{R}_{\mathrm{train}}(f, g) + \Delta_1 + \Delta_2).$$

Though seemingly simple, the accomodation of additive slack is deceptively flexible. Given probability laws P, Q on a space $\mathcal{X} = \{X \in \mathcal{X}\}$, recall their Radon-Nikodym derivative (see, e.g. Polyanskiy and Wu (2022, Chapter 2)) $\mathrm{d}\mathsf{P}(X)/\mathrm{d}\mathsf{Q}(X)$ (which may take value ∞). In the language of Radon-Nikodym derivatives, the worst-case density ratios $\nu_{x,y}$ and ν_y in Definition 2.1 can be defined as

$$\nu_{x,y} := \sup_{\mathbf{x},\mathbf{y}} \frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{x}, \mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{x}, \mathbf{y})}, \quad \nu_y := \sup_{\mathbf{y}} \frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{y})}. \tag{A.1}$$

The following corollary, whose proof we give in Appendix A.3.3, shows that we can replace the dependence on the worst-case density ratios in Lemma 3.1 with a bound that depends only on the *tails* of those density ratios:

Corollary A.1. *Recall the boundedness assumption Assm. 2.4, such that all of $|f|, |g|, |f_\star|, |g_\star|$ are uniformly at most B in magntinude. For any functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$, it holds that*

$$\mathcal{R}_{\mathsf{test}}(f, g) \le \inf_{t_1, t_2 > 0} 2(t_1 \mathcal{R}_{\mathrm{train}}[f] + t_2 \mathcal{R}_{\mathrm{train}}(f, g) + 16B^2 \Delta_{x,y}(t_1) + 16B^2 \Delta_y(t_2)),$$

where we define

$$\Delta_{x,y}(t) := \mathbb{P}_{(\mathbf{x},\mathbf{y}) \sim \mathbb{P}_{\mathrm{test}}}\left[\frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{x}, \mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{x}, \mathbf{y})} > t\right], \quad \Delta_y(t) := \mathbb{P}_{\mathbf{y} \sim \mathbb{P}_{\mathrm{test}}}\left[\frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{y})} > t\right].$$

Remark A.2. As in Section 3.3, the above bound can be refined in a function-class dependent fashion by replacing the density ratio tail-bounds in the definitions of $\Delta_{x,y}(t)$ and $\Delta_y(t)$ with the restriction of these density ratios to the σ-algebra generated by the function classes $\{(\mathbf{x}, \mathbf{y}) \mapsto f(\mathbf{x}) - f_\star(\mathbf{x}) - \beta_f(\mathbf{y}) : f \in \mathcal{F}\}$ and $\{\mathbf{y} \mapsto g(\mathbf{y}) - g_\star(\mathbf{y}) - \beta_f(\mathbf{y}) : f \in \mathcal{F}, g \in \mathcal{G}\}$. When ν_1 and ν_2, as defined in Eqs. (3.3) and (3.4), are bounded, then taking $t_1 = \nu_1$ and $t_2 = \nu_2$ recovers the bounds obtained in Section 3.3.

A.3.1. Examples with unbounded density ratios.

We give two simple examples demonstrating how the bound in Corollary A.1 can be finite even when $\nu_{x,y}$ and ν_y are not. Our first illustrative example shows that Corollary A.1 can be finite even if there are values of $\mathbf{y}$ for which the density ratios are infinte.

Example A.1 (Infinite Density Ratios)**.** Consider a discrete setting where, for simplicity, $\mathbf{x} = 1$ is deterministic, and $\mathbf{y} \in [n+1] = \{1, \dots, n+1\}$. Suppose that $\mathbf{y}$ under $\mathbb{P}_{\mathsf{train}}$ is distributed uniformly on $[n]$, and uniformly on $[n+1]$ under $\mathbb{P}_{\mathrm{test}}$. For discrete distributions, density ratios are just ratios of probabilities:

$$\frac{\mathbb{P}_{\mathrm{test}}(\mathbf{x}, \mathbf{y})}{\mathbb{P}_{\mathsf{train}}(\mathbf{x}, \mathbf{y})} = \frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{y})} = \begin{cases} \frac{n}{n+1} & \mathbf{y} \in [n] \\ \infty & \mathbf{y} = [n+1]. \end{cases} \tag{A.2}$$

Thus, the density ratios are not uniformly bounded, and may indeed take the value ∞. Still, $\mathbb{P}_{\mathbf{y} \sim \mathbb{P}_{\mathrm{test}}}[\mathbf{y} = n+1] = \frac{1}{n}$, so Corollary A.1 with $t_1 = t_2 = \frac{n}{n+1} \leq 1$yields

$$\mathcal{R}_{\mathsf{test}}(f, g) \leq 2(\mathcal{R}_{\mathrm{train}}[f] + \mathcal{R}_{\mathrm{train}}(f, g)) + \frac{64B^2}{n+1}, \tag{A.3}$$

which is not only not infinite, but indeed decays with n.

Our second example shows the sample can happen even if the density ratios are finite, but unbounded.

Example A.2. Let $\gamma_1, \gamma_2 \in (0, 1)$. Again, consider deterministic $\mathbf{x} = 1$ and discrete $\mathbf{y} \in \mathbb{Z}_{\geq 0}$ with geometric distributions. $\mathbb{P}_{\mathsf{train}}[\mathbf{y} = k] = (1-\gamma_1)\gamma_1^k, \quad \mathbb{P}_{\mathrm{test}}[\mathbf{y} = k] = (1-\gamma_2)\gamma_2^k$. Then,

$$\frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{y} = k)}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{y} = k)} = \frac{1-\gamma_2}{1-\gamma_1} \cdot \left(\frac{\gamma_2}{\gamma_1}\right)^k, \quad k \in \mathbb{Z}_{\geq 0}$$

When $\gamma_2 > \gamma_1$, $\sup_{\mathbf{y}} \frac{\mathrm{d}\,\mathbb{P}_{\mathrm{test}}(\mathbf{y})}{\mathrm{d}\,\mathbb{P}_{\mathsf{train}}(\mathbf{y})} = \infty$. Still, Corollary A.1 is non vacuous, yielding

$$\mathcal{R}_{\mathsf{test}}(f, g) \leq \inf_{t > 0} 2t(\mathcal{R}_{\mathrm{train}}[f] + \mathcal{R}_{\mathrm{train}}(f, g)) + 64B^2\, \mathbb{P}_{\mathrm{test}}\left[\frac{1-\gamma_2}{1-\gamma_1} \cdot \left(\frac{\gamma_2}{\gamma_1}\right)^{\mathbf{y}} > t\right].$$

With some algebra[1], we can compute

$$\mathcal{R}_{\mathsf{test}}(f, g) \leq \inf_{t > 0} 2t(\mathcal{R}_{\mathrm{train}}[f] + \mathcal{R}_{\mathrm{train}}(f, g)) + 64B^2 \left(t \cdot \frac{1-\gamma_1}{1-\gamma_2}\right)^{-\alpha}, \quad \alpha := \frac{\log(1/\gamma_2)}{\log \gamma_2/\gamma_1} > 0$$

An interesting feature of this example is that, even thought the density ratio in Eq. (A.4) appears to grow exponentially k, the tradeoff in t is *polynomial* due to the exponential decay of $\mathbf{y} \sim \mathbb{P}_{\mathrm{test}}$.

A.3.2. Consequences for certain f-divergences, and proof of Lemma 3.4

We show that Corollary A.1 can be instantiated for a subclass of f-divergences, which include the χ^2-divergence (with arguments ordered appropriately) as a special case. To avoid confusion with our function class f, we shall replace f with the functions $\phi : \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$.

Definition A.1 (ϕ-divergence, Chapter 2 in (Polyanskiy and Wu, 2022))**.** Given measures P, Q on the same probability space $\mathcal{X} = \{X\}$, and $\phi : \mathbb{R}_{>0} \to \mathbb{R}$, we define

$$\mathrm{D}_\phi(\mathsf{Q}, \mathsf{P}) := \int_X \phi\left(\frac{\mathrm{d}\mathsf{Q}(X)}{\mathrm{d}\mathsf{P}(X)}\right) \mathrm{d}\mathsf{P}(X)$$

where $\mathrm{d}\mathsf{Q}(X)$ and $\mathrm{d}\mathsf{P}(X)$ denote the Radon-Nikodym derivatives of Q and P, respectively, evaluated at $X \in \mathcal{X}$. By convention, it is typically required that ϕ is convex, $\phi(1) = 0$, and that $\phi(0)$ is defined via $\phi(0) = \lim_{t \to 0^+} \phi(t)$.

Observe that if the function ϕ satisfies $\phi + M$ is non-negative for some $M \in \mathbb{R}$, Markov's inequality implies

$$\mathbb{P}_{X \sim \mathsf{P}}\left[\phi\left(\frac{\mathrm{d}\mathsf{Q}(X)}{\mathrm{d}\mathsf{P}(X)}\right) > u\right] \leq \frac{\mathrm{D}_\phi(\mathsf{Q}, \mathsf{P})}{u}, \quad u \geq M.$$

[1]The missing steps are as follows. We have $\mathbb{P}_{\mathrm{test}}\left[\frac{1-\gamma_2}{1-\gamma_1} \cdot \left(\frac{\gamma_2}{\gamma_1}\right)^{\mathbf{y}} > t\right] = \mathbb{P}_{\mathrm{test}}\left[\mathbf{y} > \frac{\log(\frac{1-\gamma_1}{1-\gamma_2}) + \log t}{\log \frac{\gamma_2}{\gamma_1}}\right]$, which can be bounded using the distribution of $\mathbf{y}$ under $\mathbb{P}_{\mathrm{test}}[\mathbf{y} > u] \leq \mathbb{P}_{\mathrm{test}}[\mathbf{y} \geq u] = \mathbb{P}_{\mathrm{test}}[\mathbf{y} \geq \lceil u \rceil] \leq (\gamma_2)^u = \exp(u \log \gamma_2)$.

And, if in addition $\phi(u)$ is strictly decreasing, $\left\{\frac{\mathrm{dP}(X)}{\mathrm{dQ}(X)} > t\right\} = \left\{\frac{\mathrm{dQ}(X)}{\mathrm{dP}(X)} < \frac{1}{t}\right\} = \left\{\phi(\frac{\mathrm{dQ}(X)}{\mathrm{dP}(X)}) < \phi(\frac{1}{t})\right\}$. Thus,

$$\mathbb{P}_{X\sim\mathsf{P}}\left[\frac{\mathrm{dP}(X)}{\mathrm{dQ}(X)} > t\right] \leq \frac{\mathrm{D}_\phi(\mathsf{Q},\mathsf{P})}{\phi(1/t)}, \quad \phi(1/t) \geq M. \tag{A.4}$$

Then, Corollary A.1 directly implies the following consequence.

Corollary A.2. *Consider any non-negative and strictly decreasing functions $\phi_1, \phi_2 : \mathbb{R}_{>0} \to [-M, \infty)$; the other axioms of the ϕ-divergence need not be met. Then, for any $t_1, t_2 > 0$ for which $\phi_1(1/t_1), \phi_2(1/t_2) \geq M$,*

$$\begin{aligned}\mathcal{R}_{\mathsf{test}}(f,g) \leq\ & 2(t_1\mathcal{R}_{\mathrm{train}}[f] + t_2\mathcal{R}_{\mathrm{train}}(f,g)) \\ & + 32B^2\left(\frac{\mathrm{D}_{\phi_1}(\mathbb{P}_{\mathsf{train}}(\mathbf{x},\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{x},\mathbf{y}))}{\phi_1(1/t_1)} + \frac{\mathrm{D}_{\phi_2}(\mathbb{P}_{\mathsf{train}}(\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{y}))}{\phi_2(1/t_2)}\right),\end{aligned}$$

where $\mathrm{D}_{\phi_1}(\mathbb{P}_{\mathsf{train}}(\mathbf{x},\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{x},\mathbf{y})$ denotes the ϕ_1-divergence between the joint distribution of $(\mathbf{x},\mathbf{y})$ under $\mathsf{Q} = \mathbb{P}_{\mathsf{train}}$ and $\mathsf{P} = \mathbb{P}_{\mathrm{test}}$, and $\mathrm{D}_{\phi_2}(\mathbb{P}_{\mathsf{train}}(\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{y}))$ the ϕ_2-divergence between the marginal of $\mathbf{y}$ under these measures.

We show now describe two special cases of interest.

Example A.3 (Power Divergences)**.** A special case is when $\phi_i(t) = \frac{1}{t_i^{\alpha_i}} - 1$, $\alpha_i > 0$,[2] then for any $t_1, t_2 \geq 1$,

$$\begin{aligned}\mathcal{R}_{\mathsf{test}}(f,g) \leq\ & 2(t_1\mathcal{R}_{\mathrm{train}}[f] + t_2\mathcal{R}_{\mathrm{train}}(f,g)) \\ & + 32B^2\left(\frac{\mathrm{D}_{\phi_1}(\mathbb{P}_{\mathsf{train}}(\mathbf{x},\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{x},\mathbf{y}))}{t_1^{\alpha_1}} + \frac{\mathrm{D}_{\phi_2}(\mathbb{P}_{\mathsf{train}}(\mathbf{y}),\mathbb{P}_{\mathrm{test}}(\mathbf{y}))}{t_2^{\alpha_2}}\right).\end{aligned}$$

We obtain Lemma 3.4 as a special case:

Proof of Lemma 3.4. An archetypical example of the special case of power-divergences described above is $\phi_1(u) = \phi_2(u) = \frac{1}{u} - 1$. Then, it can be shown that $\mathrm{D}_\phi(\mathsf{Q},\mathsf{P}) = \chi^2(\mathsf{P},\mathsf{Q})$, where $\chi^2(\cdot,\cdot)$ denotes the χ^2 divergence (note the reversed order of the arguments). Thus, specializing Example A.3 further yields that, for any $t_1, t_2 \geq 1$,

$$\begin{aligned}\mathcal{R}_{\mathsf{test}}(f,g) \leq\ & 2(t_1\mathcal{R}_{\mathrm{train}}[f] + t_2\mathcal{R}_{\mathrm{train}}(f,g)) \\ & + 32B^2\left(\frac{\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{x},\mathbf{y}),\mathbb{P}_{\mathsf{train}}(\mathbf{x},\mathbf{y}))}{t_1} + \frac{\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{y}),\mathbb{P}_{\mathsf{train}}(\mathbf{y}))}{t_2}\right).\end{aligned}$$

When $\mathcal{R}_{\mathrm{train}}[f], \mathcal{R}_{\mathrm{train}}(f,g)$ are sufficiently small relative to the above $\chi^2(\cdot,\cdot)$ divergences, we can minimize over t_1, t_2 without the $t \geq 1$ constraint:

$$\mathcal{R}_{\mathsf{test}}(f,g) \leq 8B\sqrt{\mathcal{R}_{\mathrm{train}}[f]\cdot\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{x},\mathbf{y}),\mathbb{P}_{\mathsf{train}}(\mathbf{x},\mathbf{y}))} + 8B\sqrt{\mathcal{R}_{\mathrm{train}}(f,g)\cdot\chi^2(\mathbb{P}_{\mathrm{test}}(\mathbf{y}),\mathbb{P}_{\mathsf{train}}(\mathbf{y}))}.$$

□

A.3.3. Proof of Corollary A.1

We begin with a standard change-of-measure bound.

Lemma A.1 (Change of Measure)**.** *For any measurable, bounded function $h : \mathcal{X} \to [0, M]$,*

$$\mathbb{E}_{X\sim\mathsf{P}}[h(X)] \leq \inf_{t>0} M\,\mathbb{P}_{X\sim\mathsf{P}}\left[\left\{\frac{\mathrm{dP}(X)}{\mathrm{dQ}(X)} > t\right\}\right] + t\mathbb{E}_{X\sim\mathsf{Q}}[h(X)].$$

[2]Note that this ensures that $\phi_i(\cdot)$ is strictly decreasing, $\phi_i + 1 \geq 0$, as well as the f-divergence axioms $\phi_i(1) = 0$ and ϕ_i is convex.

Proof of Lemma A.1. Fix an event $\mathcal{E} \subset \mathcal{X}$, and suppose that $0 \le h(\cdot) \le M$.

$$\begin{aligned}
\mathbb{E}_{X\sim\mathsf{P}}[h(X)] &\le \mathbb{E}_{X\sim\mathsf{P}}[h(X)\mathbb{I}\{\mathcal{E}\}] + \mathbb{E}_{X\sim\mathsf{P}}[h(X)\mathbb{I}\{\mathcal{E}^c\}] \\
&\le \mathbb{E}_{X\sim\mathsf{P}}[h(X)\mathbb{I}\{\mathcal{E}\}] + M\,\mathbb{P}[\mathcal{E}^c] \\
&= M\,\mathbb{P}[\mathcal{E}^c] + \int h(X)\mathbb{I}\{\mathcal{E}\}\mathrm{d}\mathsf{P}(X) \\
&= M\,\mathbb{P}[\mathcal{E}^c] + \int h(X)\mathbb{I}\{\mathcal{E}\}\mathrm{d}\mathsf{Q}(X)\cdot\frac{\mathrm{d}\mathsf{P}(X)}{\mathrm{d}\mathsf{Q}(X)} \\
&\le M\,\mathbb{P}[\mathcal{E}^c] + \left(\int h(X)\mathbb{I}\{\mathcal{E}\}\mathrm{d}\mathsf{Q}(X)\right)\cdot\sup_{X\in\mathcal{E}}\left(\frac{\mathrm{d}\mathsf{P}(X)}{\mathrm{d}\mathsf{Q}(X)}\right) \\
&\le M\,\mathbb{P}[\mathcal{E}^c] + \mathbb{E}_{X\sim\mathsf{Q}}[h(X)]\cdot\sup_{X\in\mathcal{E}}\left(\frac{\mathrm{d}\mathsf{P}(X)}{\mathrm{d}\mathsf{Q}(X)}\right).
\end{aligned}$$

To conclulde, we take $\mathcal{E} := \{\frac{\mathrm{d}\mathsf{P}(X)}{\mathrm{d}\mathsf{Q}(X)} \le t\}$. □

Proof of Corollary A.1. We aim to establish the following:

$$\mathbb{E}_{\mathsf{test}}[(f - f_\star - \beta_f)^2] \le \nu_1\mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta_f)^2] + \Delta_1 \tag{A.5}$$
$$\mathbb{E}_{\mathsf{test}}[(g - g_\star - \beta_f)^2] \le \nu_2\mathbb{E}_{\mathsf{train}}[(g - g_\star - \beta_f)^2] + \Delta_2, \tag{A.6}$$

In view of Corollary 3.2, it suffices to check that for any $t_1, t_2 > 0$, Eqs. (A.5) and (A.6) hold with $(\nu_1, \nu_2) \leftarrow (t_1, t_2)$, and $(\Delta_1, \Delta_2) \leftarrow (16B^2\Delta_{x,y}(t_1), 16B^2\Delta_y(t_2)))$. Consider the functions of the form $h_1(\mathbf{x}, \mathbf{y}) = (f(\mathbf{x}) - f_\star(\mathbf{x}) - \beta_f(\mathbf{y}))^2$ and $h_2(\mathbf{y}) = (g(\mathbf{y}) - g_\star(\mathbf{y}) - \beta_f(\mathbf{y}))^2$. By Assm. 2.4, it holds that the image of both functions lies in $[0, 16B^2]$. The result now follows by applying Lemma A.1 with $M \leftarrow 16B^2$ to the functions h_1 (resp h_2) with $t \leftarrow t_1$ (resp. $t \leftarrow t_2$). □

A.4. Correspondence with Double ML

As we have mentioned, our results have a similar flavor to those in the literature on Neyman orthogonalization. In this section, we expand on this comparison, following the treatment in (Foster and Syrgkanis, 2019). As terminology, we refer to these idea broadly as orthogonal (statistical) learning.

The orthogonal learning setup describes a similar situation where a pair $(f_\star, g_\star)$ is unknown, but weare primarily interested in $f_\star$, referring to $g_\star$ as a nuisance function. For example, we may have $\mathbb{E}\mathbf{z} \mid \mathbf{x}, \mathbf{y} = f_\star(\mathbf{x}) + g_\star(\mathbf{y})$ as in our setup, but orthogonal learning is more general in this respect. Unlike our setting however, orthogonal learning requires some auxiliary mechanism or data to learn $\hat{g}$ such that $\mathbb{E}(\hat{g}(\mathbf{y}) - g_\star(\mathbf{y}))^2 \lesssim \mathrm{rate}_n(\mathcal{G})$. Given this initial estimate, orthogonal learning describes an algorithm and conditions under which one can learn $\hat{f}$ satisfying

$$\mathbb{E}(\hat{f}(\mathbf{x}) - f_\star(\mathbf{x}))^2 \lesssim \mathrm{rate}_n(\mathcal{F}) + \mathrm{rate}_n(\mathcal{G})^2.$$

Here $\mathrm{rate}_n(\cdot)$ should be thought of as the standard "fast" rate for learning with the function class, i.e., $\mathrm{rate}_n(\mathcal{F}) \asymp \frac{\log|\mathcal{F}|}{n}$ when $|\mathcal{F}| < \infty$. This guarantee naturally leads to a distribution shift bound of the form:

$$\mathcal{R}_{\mathsf{test}}(\hat{f}, \hat{g}) \lesssim \nu_x\left(\mathrm{rate}_n(\mathcal{F}) + \mathrm{rate}_n(\mathcal{G})^2\right) + \nu_y\mathrm{rate}_n(\mathcal{G}).$$

As with our bound, the complexity of the class $\mathcal{G}$ does not interact with distribution shifts on ν_x in a significant way. Indeed, the error rate for $\hat{f}$ has a quadratic dependence on that for $\hat{g}$, and this is typically lower order when considering distribution shift settings. Thus, at a conceptual level, orthogonal learning can provide a similar robustness to heterogenous distribution shifts as our results.

Quantitatively, the bound should be compared with our Theorem 3. The general takeaway is that our bound is worse in at least two respects. First, the quadratic dependence on the rate for $\mathcal{G}$ cannot exploit localization in our setup, so our bound is weaker when $\mathcal{G}$ is small. This is why we need hypercontractivity conditions to obtain favorable rates when $\mathcal{G}$ is finite/parametric, which is not required for orthogonal learning. Second, when considering distribution shift, we incur a dependence on $\nu_{x,y}$ rather than just ν_x. This arises from the identifiability issues that are inherent in our setting, which can be resolved in the orthogonal learning setting due to the auxiliary mechanism for estimating $g_\star$.

On the other hand, our results compare favorable at a qualitative level. Most importantly, our bound applies to ERM directly while orthogonal learning requires algorithmic modifications, which in turn require more modeling of the data generating process. Additionally, while we do not believe the assumptions are formally comparable, we view ours as somewhat more practical. Specifically, it is rather uncommon that auxiliary information for estimating the nuisance parameter is available; yet in canonical settings for orthogonal learning, we can show that conditional completeness holds. One such example of the latter is the PLR model above.

B. Proof of Main Technical Results

This section provides the proofs of the most significant technical results in the paper. Specifically, Appendix B.1 give the proofs of the excess error decompositions, Lemma 3.1. Appendix A.2 establishes Lemmas 3.2 and 3.3, which refine upper bounds bounds on the distribution-shift term ν_1. Next, we prove Proposition 4.2 which upper bounds $\mathcal{R}_{\text{train}}[f, g]$, and thus, in view of Lemma 3.1, $\mathcal{R}_{\text{train}}[g; f]$. Appendix B.3 gives the proof of Proposition 4.3 which provides refined control of $\mathcal{R}_{\text{train}}[f]$; the key step is an (empirical) excess-risk decomposition, Lemma B.3, which we prove in Appendix B.5. Finally, Appendix B.6 establishes our Hölder inequality for Rademacher complexities of Hadamard product classes (Proposition 4.4). Lastly, Appendix B.7 derives the standard Dudley integral bound from the aforementioned proposition for product classes.

B.1. Proof of Lemma 3.1

Write $\beta = \beta_f$ for simplicity. For any environment e and any triplet (f, g, β), the polarization identity yields

$$\begin{aligned}
\mathcal{R}_{\mathsf{e}}(f, g) &= \mathbb{E}_{\mathsf{e}}[(f + g - f_\star - g_\star)^2] \\
&= \mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta + g - g_\star + \beta)^2] \\
&= \mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)^2] + \mathbb{E}_{\mathsf{e}}[(g - g_\star + \beta)^2] + 2\mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)(g - g_\star + \beta)] \\
&= \mathcal{R}_{\mathsf{e}}[f] + \mathcal{R}_{\mathsf{e}}[g; f] + 2\mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)(g - g_\star + \beta)].
\end{aligned}$$

For any e (in particular, $\mathsf{e} = \mathsf{test}$), we have

$$\begin{aligned}
\mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)(g - g_\star + \beta)] &\le \mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)^2]^{1/2} \cdot \mathbb{E}_{\mathsf{e}}[(g - g_\star + \beta)^2]^{1/2} \\
&\le \mathbb{E}_{\mathsf{e}}[(f - f_\star - \beta)^2] + \mathbb{E}_{\mathsf{e}}[(g - g_\star + \beta)^2] = \mathcal{R}_{\mathsf{e}}[f] + \mathcal{R}_{\mathsf{e}}[g; f].
\end{aligned}$$

Hence,

$$\mathcal{R}_{\mathsf{e}}(f, g) \le 2(\mathcal{R}_{\mathsf{e}}[f] + \mathcal{R}_{\mathsf{e}}[g; f]).$$

When $\mathsf{e} = \mathsf{train}$, the fact that $\beta(\mathbf{y}) = \mathbb{E}_{\mathsf{train}}[(f - f_\star)(\mathbf{x}) \mid \mathbf{y}]$ implies

$$\begin{aligned}
\mathbb{E}_{\mathsf{train}}[(f - f_\star - \beta)(g - g_\star + \beta)] &= \mathbb{E}_{\mathsf{train}}[((f - f_\star)(\mathbf{x}) - \beta(\mathbf{y})) \cdot (g - g_\star + \beta)(\mathbf{y}) \mid \mathbf{y}] \\
&= \mathbb{E}_{\mathsf{train}}[(\mathbb{E}_{\mathsf{train}}[(f - f_\star)(\mathbf{x}) \mid \mathbf{y}] - \beta(\mathbf{y})) \cdot (g - g_\star + \beta)(\mathbf{y})] = 0.
\end{aligned}$$

Thus,

$$\mathcal{R}_{\text{train}}(f, g) = \mathcal{R}_{\text{train}}[f] + \mathcal{R}_{\text{train}}[g; f].$$

This proves the first two parts of the lemma. For the last part, we have

$$\begin{aligned}
\mathcal{R}_{\mathsf{test}}(f, g) &\le 2\mathcal{R}_{\mathsf{test}}[f] + 2\mathcal{R}_{\mathsf{test}}[g; f] && \text{(Lemma 3.1)} \\
&\overset{(i)}{\le} 2\nu_{x,y}\mathcal{R}_{\text{train}}[f] + 2\nu_y\mathcal{R}_{\text{train}}[g; f] && \text{(B.1)} \\
&\overset{(ii)}{\le} 2\nu_{x,y}\mathcal{R}_{\text{train}}[f] + 2\nu_y\mathcal{R}_{\text{train}}(f, g),
\end{aligned}$$

where (i) invokes Definition 2.1, and where in (ii), $\mathcal{R}_{\text{train}}[f] \ge 0$ and $\mathcal{R}_{\text{train}}(f, g) = \mathcal{R}_{\text{train}}[f] + \mathcal{R}_{\text{train}}[g; f]$ implies $\mathcal{R}_{\text{train}}[g; f] \le \mathcal{R}_{\text{train}}(f, g)$. □

B.2. Proof of Proposition 4.2

This section proves Proposition 4.2, which we use to upper bound $\mathcal{R}_{\mathrm{train}}[f, g]$, and thus, by way of Lemma 3.1, $\mathcal{R}_{\mathrm{train}}[g; f]$. We begin by establishing the following more or less standard guarantee (see, e.g. (Liang et al., 2015)) for a generic function class $\mathcal{H}$. which controls the so-called "basic inequality" in square-loss learning (see, e.g. Wainwright (2019, Chapter 13)).

Lemma B.1. *Let $\mathcal{H}$ be a functions from $\mathcal{W} \to [-B, B]$ containing the zero function $h_0(w) \equiv 0$. Fix $\sigma > 0$, $\tau \geq 1$. Then, for any probability measure P, $\mathbf{w}_1, \dots, \mathbf{w}_n \overset{\text{i.i.d}}{\sim} \mathbb{P}$ and i.i.d. standard normal random variables $\boldsymbol{\xi}_1 \dots, \boldsymbol{\xi}_n$, the following holds with probability $1 - \delta$*

$$\sup_{h \in \mathcal{H}} \frac{1}{4}\|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n} + \frac{2\tau\sigma}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h(\mathbf{w}_i) \lesssim \gamma_{n,\delta,\sigma}(\mathcal{H})^2,$$

where $C > 0$ is a universal constant and

$$\gamma_{n,\sigma,\tau}(\mathcal{H}, \delta)^2 = \delta_{n,\mathscr{R}}(\mathcal{H}, B)^2 + \tau^2 \delta_{n,\mathscr{G}}(\mathcal{H}, \sigma)^2 + \frac{(\tau^2\sigma^2 + B^2)\log(1/\delta)}{n}.$$

Proof. We have

$$\begin{aligned}
&\frac{1}{4}\|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n} + \frac{2\tau\sigma}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h(\mathbf{w}_i) \\
&= \frac{1}{4}\underbrace{\left(\|h\|^2_{\mathcal{L}_2(P)} - 2\|h(\mathbf{w}_{1:n})\|^2_{2,n}\right)}_{\mathrm{Term}_1} + \underbrace{\left(\frac{-1}{2}\|h(\mathbf{w}_{1:n})\|^2_{2,n} + \frac{2\tau\sigma}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h(\mathbf{w}_i)\right)}_{\mathrm{Term}_2}
\end{aligned}$$

By Lemma C.14 and Lemma C.10, respecitively, the following holds with probability at least $1 - \delta$,

$$\begin{aligned}
\mathrm{Term}_1 &\lesssim \left(\delta_{n,\mathscr{R}}(\mathcal{H}, B)^2 + \frac{B^2\log(1/\delta)}{n}\right) \\
\mathrm{Term}_2 &\lesssim \tau^2\left(\frac{\sigma^2\log(1/\delta)}{n} + \delta_{n,\mathscr{G}}(\mathcal{H}, \sigma)^2\right).
\end{aligned}$$

Summing concludes. □

Proof of Proposition 4.2. Let $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i)$ and $\mathbf{w} = (\mathbf{x}, \mathbf{y})$, and set $h_\star := f_\star + g_\star$ and $\hat{h}_n := \hat{f}_n + \hat{g}_n$. Note that $\hat{h}_n - h_\star \in \mathcal{H}_{\mathtt{cnt}}$. It follows from the so-called "basic ineqality" Wainwright (2019, Eq. 13.36) that

$$\|(\hat{h}_n - h_\star)(\mathbf{w}_{1:n})\|^2_{2,n} - 2\boldsymbol{\xi}_i(\hat{h}_n - h_\star)(\mathbf{w}_i) \leq 0,$$

Thus, by adding and subtracting $\frac{1}{4}\mathbb{E}[\hat{h}_n(\mathbf{w})^2]$ and rearranging

$$\mathbb{E}[(\hat{h}_n - h_\star)(\mathbf{w})^2] \leq 4\left(-\mathbb{E}[(\hat{h}_n - h_\star)(\mathbf{w})^2] + \|(\hat{h}_n - h_\star)(\mathbf{w}_{1:n})\|^2_{2,n} + 2\boldsymbol{\xi}_i(\hat{h}_n - h_\star)(\mathbf{w}_i)\right).$$

Passing to the supremum over all $h \in \mathcal{H}_{\mathtt{cnt}} := \mathcal{F} + \mathcal{G} - (f_\star + g_\star)$ and invoking Lemma B.1 shows that

$$\mathcal{R}_{\mathrm{train}}(\hat{f}_n, \hat{g}_n) \lesssim \gamma_n(\delta)^2 := \delta_{n,\mathscr{R}}(\mathcal{H}_{\mathtt{cnt}}, B)^2 + \delta_{n,\mathscr{G}}(\mathcal{H}_{\mathtt{cnt}}, \sigma)^2 + \frac{(B^2 + \sigma^2)\log(1/\delta)}{n}.$$

□

B.3. Proof of Proposition 4.3

This section proves Proposition 4.3, which we use to bound $\mathcal{R}_{\text{train}}[f]$. It is considerably more involved than the proof of Proposition 4.2, as we need to argue that the "large" class $\mathcal{G}$ does not heavily obfuscate recovery in the class $\mathcal{F}$. Throughout, let us use $\mathbf{w} = (\mathbf{x}, \mathbf{y})$, $\mathbf{w}_i = (\mathbf{x}_i, \mathbf{y}_i)$, and $\mathcal{W} = \mathcal{X} \times \mathcal{Y}$. Recall the sets

$$\mathcal{F}_{\texttt{cnt}} := \{f - \beta_f - f_\star : f \in \mathcal{F}\}, \quad \mathcal{G}_{\texttt{cnt}} := \{g + \beta_f - f_\star : f \in \mathcal{F}, g \in \mathcal{G}\}.$$

We begin by uniformly bounding the elements of $\mathcal{F}_{\texttt{cnt}}$ and $\mathcal{G}_{\texttt{cnt}}$.

Lemma B.2. *For any $h \in \mathcal{F}_{\texttt{cnt}} \cup \mathcal{G}_{\texttt{cnt}}$, $\sup_{w \in \mathcal{H}} |h(w)| \le 4B$, where B is as in Assm. 2.4.*

Proof. Follows directly from Assm. 2.4 and the fact that $\beta_f(y) := \mathbb{E}[(f - f_\star)(\mathbf{x}) \mid \mathbf{y} = y]$. □

In Appendix B.5, we prove the following, which shows that the conditional completeness allows us to decompose $\mathcal{R}_{\text{train}}[\hat{f}_n]$ across these two terms, the first of which represents a standard excess risk in terms of $\mathcal{F}$, and the latter of which measures the contamination due to errors in $\mathcal{G}$. This bound can be thought of as a careful refinement of the standard "basic inequality" (Wainwright, 2019, Eq. 13.36).

Lemma B.3. *If $(\mathcal{F}, \mathcal{G})$ satisfies γ-conditional completeness, then for any empirical risk minimizer $(\hat{f}_n, \hat{g}_n)$ for which $\mathcal{R}_{\text{train}}(\hat{f}_n, \hat{g}_n) \le \gamma^2$, the following bound holds deterministically:*

$$\mathcal{R}_{\text{train}}[\hat{f}_n] \le 8 \sup_{h_1 \in \mathcal{F}_{\texttt{cnt}}} \left(\frac{1}{4}\mathbb{E}[h_1(\mathbf{w})^2] - \frac{1}{2}\|h_1(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma \cdot \frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_1(\mathbf{w}_i) \right) \tag{Term$_1$}$$

$$+ 32 \sup_{h_1 \in \mathcal{F}_{\texttt{cnt}}, h_2 \in \mathcal{G}_{\texttt{cnt}}(\gamma)} \left(-\frac{1}{2}\mathbb{E}[h_1(\mathbf{w})^2] - \frac{4}{n}\sum_{i=1}^n h_1(\mathbf{w}_i) \cdot h_2(\mathbf{w}_i) \right). \tag{Term$_2$}$$

In the remainder of the proof, we apply various learning-theoretic tools to upper bound the right-hand side of Lemma B.3. These tools, and their proofs, are detailed in Appendix C. While the tools themselves are more-or-less standard, deriving from the offset-Rademacher arguments in (Liang et al., 2015), their application to the refined decomposition in Lemma B.3 yields the novelty of Proposition 4.3.

Proof of Proposition 4.3. We prove the variant of the lemma of the lemma involving $\delta_{n,\text{cross}}$, and explain how to modify the proof to obtain dependence on $\bar{\delta}_{n,\text{cross}}$ at the end.

The first term in the above display can be bounded directly from Lemma B.1 with τ set to 1, yielding

$$\text{Term}_1 \lesssim \delta_{n,\mathscr{R}}(\mathcal{F}_{\texttt{cnt}}, B)^2 + \delta_{n,\mathscr{G}}(\mathcal{F}_{\texttt{cnt}}, \sigma)^2 + \frac{(\sigma^2 + B^2)\log(1/\delta)}{n}. \tag{B.2}$$

To bound the second term, we use a localization argument. Define the doubly-localized term

$$\Psi(r, \gamma) := \sup_{h_1 \in \mathcal{F}_{\texttt{cnt}}(r), h_2 \in \mathcal{G}_{\texttt{cnt}}(\gamma)} \left(-\frac{r^2}{2} - \frac{4}{n}\sum_{i=1}^n h_1(\mathbf{w}_i) \cdot h_2(\mathbf{w}_i) \right).$$

Following the same argument as in Lemma C.9, one can check that

$$\text{Term}_2 = 32 \sup_{r > 0} \Psi(r, \gamma),$$

and that

$$\text{Term}_2 \le 32 \cdot \inf \left\{ r^2 : \Psi(r, \gamma) \le \frac{r^2}{2} \right\}. \tag{B.3}$$

Hence, let us exhibit an r for which $\Psi(r,\gamma) \le 0$ with high probability. Define the shorthand $\mathcal{H}_{r,\gamma} := \mathcal{F}_{\mathtt{cnt}}(r) \odot \mathcal{G}_{\mathtt{cnt}}(\gamma)$. For an $r > 0$ to be chosen, Lemma C.8 implies that with probability $1 - \delta/4$,

$$\Psi(r,\gamma) \le c\left(\mathbb{E}_{\mathbf{w}_{1:n}}[\mathscr{R}_n(\mathcal{H}_{r,\gamma}[\mathbf{w}_{1:n}])] + r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{B\log(1/\delta)}{n}\right) - \frac{r^2}{8}$$

where $c \ge 1$ is a universal constant. By AM-GM, we have

$$\begin{aligned}\Psi(r,\gamma) &\le c\left(\mathbb{E}_{\mathbf{w}_{1:n}}[\mathscr{R}_n(\mathcal{H}_{r,\gamma}[\mathbf{w}_{1:n}])] + \frac{r^2}{16c} + \frac{(8c+B)\log(1/\delta)}{n}\right) - \frac{r^2}{8} \\ &= c\left(\mathbb{E}_{\mathbf{w}_{1:n}}[\mathscr{R}_n(\mathcal{H}[\mathbf{w}_{1:n}])] - \frac{r^2}{16c} + \frac{(8c+B)\log(1/\delta)}{n}\right) \qquad \text{(B.4)}\end{aligned}$$

We now compute

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}_{1:n}\sim P}[\mathscr{R}_n(\mathcal{H}_{r,\gamma}[\mathbf{w}_{1:n}])] - \frac{r^2}{16c} \\ &= \mathbb{E}_{\mathbf{w}_{1:n}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h\in\mathcal{H}_{r,\gamma}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h(\mathbf{w}_i)\right] - \frac{r^2}{16c} \\ &= \mathbb{E}_{\mathbf{w}_{1:n}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}(r), h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(\mathbf{w}_i)h_2(\mathbf{w}_i)\right] - \frac{r^2}{16c} && \text{(Definition of } \mathcal{H}_{r,\gamma}) \\ &= \mathbb{E}_{\mathbf{w}_{1:n}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}(r), h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(\mathbf{w}_i)h_2(\mathbf{w}_i) - \frac{1}{16c}\mathbb{E}[h_1(\mathbf{w})^2]\right] && \text{(Localization of } \mathcal{F}_{\mathtt{cnt}}(r)) \\ &\le \frac{1}{16c}T_1 + \frac{1}{32c}T_2,\end{aligned}$$

where we define

$$\begin{aligned}T_1 &:= \mathbb{E}_{\mathbf{w}_{1:n}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}(r), h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{16c}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(\mathbf{w}_i)h_2(\mathbf{w}_i) - \frac{1}{2}\|h_1(\mathbf{w}_{1:n})\|_{2,n}^2\right] \\ T_2 &:= \mathbb{E}_{\mathbf{w}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}} \|h_1(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\mathbb{E}[h_1(\mathbf{w})^2]\right].\end{aligned}$$

Bounding T_1. To being, we remove the localization of $\mathcal{F}_{\mathtt{cnt}}$ in the term T_1, upper bounding

$$T_1 \le T_1' := \mathbb{E}_{\mathbf{w}_{1:n}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}, h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{16c}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(\mathbf{w}_i)h_2(\mathbf{w}_i) - \frac{1}{2}\|h_1(\mathbf{w}_{1:n})\|_{2,n}^2\right]$$

For any fixed $w_{1:n} \in \mathcal{W}^n$, consider the process

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(w_i)h_2(w_i).$$

Introduce the class

$$\tilde{\mathcal{H}}[\rho] := \{h_1 \cdot h_2 : h_1 \in \mathcal{F}_{\mathtt{cnt}}, h_2 \in \mathcal{G}_{\mathtt{cnt}}(\gamma), \|h_1(w_{1:n})\|_{2,n} \le \rho\}, \qquad \text{(B.5)}$$

which localizes h_1 at empirical $\mathcal{L}_2$-norm ρ. Note that that, for any $\tilde{h} \in \tilde{\mathcal{H}}[\rho]$, we can write $\tilde{h} = h_1 \cdot h_2$ where

$$\|\tilde{h}\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^n h_1(w_i)^2 h_2(w_i)^2 \le \frac{4B^2}{n}\sum_{i=1}^n h_1(w_i)^2 = 4B^2\|h_1(w_{1:n})\|_{2,n}^2 \le 4B^2\rho^2,$$

where the first inequality is by Lemma B.2 and second by definition of $\tilde{\mathcal{H}}[\rho]$. Again by Lemma B.2, $\max_i \sup_{\tilde{h}\in\tilde{\mathcal{H}}[\rho]} \le 4B^2$. It follows by Lemma C.7 that the following holds with probability $1-\delta$

$$\begin{aligned}\mathbf{Z}[\rho, w_{1:n}] &:= \sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}:\|h_1(w_{1:n})\|_{2,n}\le\rho} \sup_{h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i \tilde{h}(w_i) \\ &= \sup_{\tilde{\mathcal{H}}[\rho]} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i \tilde{h}(w_i) \\ &\lesssim \mathbb{E}\left[\sup_{\tilde{\mathcal{H}}[\rho]} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i \tilde{h}(w_i)\right] + B\rho\sqrt{\log(1/\delta)/n} + \frac{B^2}{n}.\end{aligned}$$

Applying Lemma C.9 with the classes

$$\mathbb{V} = \mathcal{H}[w_{1:n}], \quad \mathbf{u}_i = (\boldsymbol{\varepsilon}_i, i), \quad \Phi := \{\mathbf{u}_i \mapsto \boldsymbol{\varepsilon}_i h_2(w_i) : h_2 \in \mathcal{G}_{\mathtt{cnt}}(\gamma)\}$$

and constants

$$c_1 \lesssim 1, \quad c_2 \lesssim B, \quad c_2 \lesssim B^2, \sigma = 1, \quad \tau \lesssim 1,$$

we conclude

$$T_1' \le \sup_{w_{1:n}} \mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}, h_2\in\mathcal{G}_{\mathtt{cnt}}(\gamma)} \frac{16c}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h_1(w_i)h_2(w_i) - \frac{1}{2}\|h_1(w_i)\|_{2,n}^2\right] \lesssim \delta_n^2 + \frac{B^2}{n}, \tag{B.6}$$

where

$$\delta_n^2 := \inf\left\{\rho^2 : \sup_{w_{1:n}} \mathbb{E}[\mathbf{Z}[\rho, w_{1:n}]] \le \frac{\rho^2}{2}\right\}.$$

Note that

$$\sup_{w_{1:n}} \mathbb{E}[\mathbf{Z}[\rho, w_{1:n}]] := \sup_{w_{1:n}} \mathscr{R}_n(\mathcal{F}_{\mathtt{cnt}}[\rho, w_{1:n}] \odot \mathcal{G}_{\mathtt{cnt}}(\gamma)[w_{1:n}]),$$

so that

$$\begin{aligned}\delta_n^2 &= \inf\{\rho^2 : \sup_{w_{1:n}} \mathscr{R}_n(\mathcal{F}_{\mathtt{cnt}}[\rho, w_{1:n}] \odot \mathcal{G}_{\mathtt{cnt}}(\gamma)[w_{1:n}]) \le \frac{\rho^2}{2}\} \\ &:= \delta_{n,\mathrm{cross}}^2(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma)), \qquad \text{(Definition 4.5)}\end{aligned}$$

so that by Eq. (B.6),

$$T_1 \le T_1' \lesssim \delta_{n,\mathrm{cross}}^2(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma)) + \frac{B^2}{n}. \tag{B.7}$$

Bounding T_2. This second term can be bounded by Lemma C.13 and is at most

$$T_2 \lesssim \frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, 4B)^2 \lesssim \frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 \tag{B.8}$$

where the first inequality also uses Lemma B.2 to bound $\sup_w |h(w)| \le 4B$ for $h \in \mathcal{F}_{\mathtt{cnt}}(r)$, and the second uses Lemma C.3 to remove the factor of 4. Hence, with probability $1-\delta/4$, the following inequality holds for any fixed $r > 0$:

Concluding the proof Combining Eqs. (B.4), (B.7) and (B.8) gives that with probability $1-\delta$,

$$\begin{aligned}\Psi(r,\gamma) &\lesssim T_1 + T_2 + \frac{(1+B)\log(1/\delta)}{n} \\ &\lesssim \frac{B^2 + (1+B)\log(1/\delta)}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta^2_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma)) \\ &\lesssim \frac{(1+B^2)\log(1/\delta)}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta^2_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma)).\end{aligned}$$

Hence, if for a sufficiently large constant c', we take

$$r^2 := c'\left(\frac{(1+B^2)\log(1/\delta)}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta^2_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma))\right),$$

then $\Psi(r,\gamma) \le \frac{r^2}{2}$ with probability at least $1-\delta$. Therefore by Eq. (B.3), we conclude that with probability $1-\delta$,

$$\mathrm{Term}_2 \lesssim \frac{(1+B^2)\log(1/\delta)}{n} + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta^2_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma)).$$

Combining with the bound on Term_1 due to Eq. (B.2) concludes the proof.

B.4. A modification of Proposition 4.3

For sharper rates with finite function classes (Appendix D), we modify Proposition 4.3 as follows.

Definition B.1 (Population-Localized Cross-Critical Radius)**.**

$$\bar{\delta}_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{H}) := \inf\left\{ r : \mathbb{E}_{\mathbf{w}_{1:n}} \mathscr{R}_n((\mathcal{F}_{\mathtt{cnt}}(r) \odot \mathcal{H})[\mathbf{w}_{1:n}]) \le \frac{r^2}{2} \right\}. \tag{B.9}$$

Proposition B.4. *Suppose that $(\mathcal{F}, \mathcal{G})$ satisfy γ-conditional completeness. Then, whenever $\mathcal{R}_{\mathsf{train}}[\hat{g}_n; \hat{f}_n] \le \gamma$, the following holds with probability at least $1-\delta$,*

$$\mathcal{R}_{\mathsf{train}}[\hat{f}_n] \lesssim \bar{\delta}^2_{n,\mathrm{cross}}(\mathcal{F}_{\mathtt{cnt}}; \mathcal{G}_{\mathtt{cnt}}(\gamma))^2 + \delta_{n,\mathscr{R}}(\mathcal{F}_{\mathtt{cnt}}, B)^2 + \delta_{n,\mathscr{G}}(\mathcal{F}_{\mathtt{cnt}}, \sigma)^2 + \frac{(\sigma^2 + B^2)\log(1/\delta)}{n}.$$

Modification to obtain dependence on $\bar{\delta}_{n,\mathrm{cross}}$. To obtain a dependence on $\bar{\delta}_{n,\mathrm{cross}}$, we change our bound the term T_1 above to use Lemma C.12 instead of Lemma C.9. The details are very similar. □

B.5. Proof of Lemma B.3

This section establishes the generalized excess-risk decomposition which forms the basis of the argument in the previous section, and which decouples - via conditional-completeness - the recovery of $f \in \mathcal{F}$ with conflation by $g \in \mathcal{G}$.

Recall $\mathbf{w} = (\mathbf{x}, \mathbf{y})$ and for $f \in \mathcal{F}$, define

$$h_f(\mathbf{w}) := (f - f_\star)(\mathbf{x}) - \beta_f(\mathbf{y}),$$

and note that $h_f \in \mathcal{F}_{\texttt{cnt}}$. Then, for any $f \in \mathcal{F}$, $g \in \mathcal{F}$, and $g_0 \in \mathcal{G}$, we have

$$\begin{aligned}
&\hat{\mathcal{L}}_n(f,g) - \hat{\mathcal{L}}_n(f_\star, g_0) \\
&= \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i) + g_0(\mathbf{y}_i) - \mathbf{z}_i)^2 - (f_\star(\mathbf{x}_i) + g_0(\mathbf{y}_i) - \mathbf{z}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((f-f_\star)(\mathbf{x}_i) - \sigma\boldsymbol{\xi}_i + (g-g_\star)(\mathbf{y}_i))^2 - (-\sigma\boldsymbol{\xi}_i + (g_0-g_\star)(\mathbf{y}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((f-f_\star)(\mathbf{x}_i) - \beta_f(\mathbf{y}_i) - \sigma\boldsymbol{\xi}_i + (g-g_\star+\beta_f)(\mathbf{y}_i))^2 - (-\sigma\boldsymbol{\xi}_i + (g_0-g_\star)(\mathbf{y}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((f-f_\star)(\mathbf{x}_i) - \beta(\mathbf{y}_i) - \sigma\boldsymbol{\xi}_i + (g-g_\star+\beta_f)(\mathbf{y}_i))^2 - (-\sigma\boldsymbol{\xi}_i + (g-g_\star+\beta_f)(\mathbf{y}_i))^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^n (-\sigma\boldsymbol{\xi}_i + (g-g_\star+\beta_f)(\mathbf{y}_i))^2 - (-\sigma\boldsymbol{\xi}_i + (g_0-g_\star)(\mathbf{y}_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^n ((f-f_\star)(\mathbf{x}_i) - \beta_f(\mathbf{y}_i))^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i((f-f_\star)(\mathbf{x}_i) - \beta_f(\mathbf{y}_i)) \\
&\qquad + \frac{2}{n}\sum_{i=1}^n ((f-f_\star)(\mathbf{x}_i) - \beta_f(\mathbf{y}_i))\cdot(g-g_\star+\beta_f)(\mathbf{y}_i)) \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^n (-\sigma\boldsymbol{\xi}_i + (g-g_\star+\beta_f)(\mathbf{y}_i))^2 - (-\sigma\boldsymbol{\xi}_i + (g_0-g_\star)(\mathbf{y}_i))^2}_{\mathrm{Remainder}(g_0; f, g)}
\end{aligned}$$

Applying the definition of h_f, the above admits the more compact form

$$\begin{aligned}
\hat{\mathcal{L}}_n(f,g) - \hat{\mathcal{L}}_n(f_\star, g_0) := \frac{1}{n}\|h_f(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_f(\mathbf{w}_i) + \frac{2}{n}\sum_{i=1}^n h_f(\mathbf{w}_i)\cdot(g-g_\star+\beta_f)(\mathbf{y}_i)) \\
+ \mathrm{Remainder}(g_0; f, g).
\end{aligned}$$

By γ-conditional completeness, we have that if $\mathcal{R}_{\mathrm{train}}[f,g] \le \gamma^2$, then we may select $\tilde{g} = g + \beta_f \in \mathcal{G}$ so that $\mathrm{Remainder}(g_0; f, g) = 0$. Similarly, if we now consider $\hat{f}_n, \hat{g}_n$ to be empirical risk minimizers of $\hat{\mathcal{L}}_n(f,g)$, it must hold that $\hat{\mathcal{L}}_n(\hat{f}_n, \hat{g}_n) - \inf_{g_0\in\mathcal{G}} \hat{\mathcal{L}}_n(f_\star, g_0) \le 0$. Thus,

$$\begin{aligned}
0 &\ge \hat{\mathcal{L}}_n(f,g) - \inf_{g_0\in\mathcal{G}} \hat{\mathcal{L}}_n(f_\star, g_0) \\
&\ge \hat{\mathcal{L}}_n(f,g) - \hat{\mathcal{L}}_n(f_\star, \tilde{g}) \\
&\ge \frac{1}{n}\|h_{\hat{f}_n}(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_{\hat{f}_n}(\mathbf{w}_i) + \frac{2}{n}\sum_{i=1}^n h_{\hat{f}_n}(\mathbf{w}_i)\cdot(g-g_\star+\beta_{\hat{f}_n})(\mathbf{y}_i)) \\
&= \frac{1}{n}\|h_{\hat{f}_n}(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_{\hat{f}_n}(\mathbf{w}_i) + \frac{2}{n}\sum_{i=1}^n h_{\hat{f}_n}(\mathbf{w}_i)\cdot(g-g_\star+\beta_{\hat{f}_n})(\mathbf{y}_i))
\end{aligned}$$

Note that $\mathbb{E}[h_{\hat{f}_n}(\mathbf{w})^2]$ is precisely equal to $\mathcal{R}_{\mathrm{train}}[\hat{f}_n]$. Adding and substracting an η multiple of this term for η tunable,

$$\begin{aligned}
\eta\mathcal{R}_{\mathrm{train}}[\hat{f}_n] \ge \frac{1}{n}\|h_{\hat{f}_n}(\mathbf{w}_{1:n})\|_{2,n}^2 - \eta\mathbb{E}[h_{\hat{f}_n}(\mathbf{w})^2] - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_{\hat{f}_n}(\mathbf{w}_i) \\
+ \frac{2}{n}\sum_{i=1}^n h_{\hat{f}_n}(\mathbf{w}_i)\cdot(g-g_\star+\beta_{\hat{f}_n})(\mathbf{y}_i)).
\end{aligned}$$

Rearranging,

$$\begin{aligned}\mathcal{R}_{\mathrm{train}}[\hat{f}_n] &\le \eta^{-1}\left(\eta\mathbb{E}[h_{\hat{f}_n}(\mathbf{w})^2] - \frac{1}{n}\|h_{\hat{f}_n}(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_{\hat{f}_n}(\mathbf{w}_i)\right)\\ &\quad + \eta^{-1}\left(-\frac{1}{2n}\sum_{i=1}^n h_{\hat{f}_n}(\mathbf{w}_i)\cdot(g - g_\star + \beta_{\hat{f}_n})(\mathbf{y}_i))\right)\\ &= \eta^{-1}\left(2\eta\mathbb{E}[h_{\hat{f}_n}(\mathbf{w})^2] - \frac{1}{n}\|h_{\hat{f}_n}(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_{\hat{f}_n}(\mathbf{w}_i)\right)\\ &\quad + \eta^{-1}\left(-\frac{\eta}{n}\mathbb{E}[h_{\hat{f}_n}(\mathbf{w})^2] - \frac{1}{2n}\sum_{i=1}^n h_{\hat{f}_n}(\mathbf{w}_i)\cdot(g - g_\star + \beta_{\hat{f}_n})(\mathbf{y}_i))\right).\end{aligned}$$

Finally, note that $h_{\hat{f}_n} \in \mathcal{F}_{\mathtt{cnt}}$. As established above, $g - \beta_{\hat{f}_n} \in \mathcal{G}$ by conditional completeness, so $g - g_\star + \beta_{\hat{f}_n} \in \mathcal{G}_{\mathtt{cnt}}$. In fact, the condition $\mathcal{R}_{\mathrm{train}}(\hat{f}_n, \hat{g}_n) \le \gamma^2$ implies via Lemma 3.1 that $\mathbb{E}[(g - \beta_{\hat{f}_n} - g_\star)^2] \le \gamma^2$, so that $g - g_\star + \beta_{\hat{f}_n} \in \mathcal{G}_{\mathtt{cnt}}(\gamma)$. Thus, we may pass to a supremum on the right-hand side equations:

$$\begin{aligned}\mathcal{R}_{\mathrm{train}}[\hat{f}_n] &\le \eta^{-1}\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}}\left(2\eta\mathbb{E}[h_1(\mathbf{w})^2] - \frac{1}{n}\|h_1(\mathbf{w}_{1:n})\|_{2,n}^2 - 2\sigma\cdot\frac{1}{n}\sum_{i=1}^n \boldsymbol{\xi}_i h_1(\mathbf{w}_i)\right)\\ &\quad + \eta^{-1}\sup_{h_1\in\mathcal{F}_{\mathtt{cnt}}, h_2\in\mathcal{G}_{\mathtt{cnt}}}\left(-\frac{\eta}{n}\mathbb{E}[h_1(\mathbf{w})^2] - \frac{1}{2n}\sum_{i=1}^n h_1(\mathbf{w}_i)\cdot h_2(\mathbf{w}_i)\right).\end{aligned}$$

Selecting $\eta = 1/8$ conclues.

B.6. Proof of Proposition 4.4

This section establishes the Hölder-style inequality for Rademacher complexities of product classes. For completeness, we begin by reproducing a standard bound on the Rademacher complexity of finite function classes.

Lemma B.5. *Let $\mathbb{V} \subset \mathbb{R}^n$ be a finite set. Then,*

$$\mathscr{R}_n(\mathbb{V}) \le \mathsf{rad}_2(\mathbb{V})\min\{1, \sqrt{2\log|\mathbb{V}|/n}\}$$

The same bounds also hold for $\hat{\mathscr{G}}_n, \mathscr{G}_n$, and more generally, whenever the variables $\boldsymbol{\varepsilon}_i$ in the definition of the Rademacher complexities are replaced by arbitrary 1-subGaussian variables.[3]

Proof. Let us bound $\hat{\mathscr{R}}_n$, with $\boldsymbol{\varepsilon}_i$ replaced by arbitrary 1-subGaussian random variables. Recall the definition of a 1-subGaussian variable $\boldsymbol{\varepsilon}$: $\log\mathbb{E}[\exp(\lambda\boldsymbol{\varepsilon})] \le \frac{1}{2}\lambda^2$ (it is standard that Gaussian random variables and Rademacher variables satisfy this inequality). By Taylor expanding $\log\mathbb{E}[\exp(\lambda\boldsymbol{\varepsilon})] = \log(1 + \sum_{i\ge1}(\lambda\mathbb{E}[\boldsymbol{\varepsilon}^i]/i!)$, it follows that $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$, and $\mathbb{E}[\boldsymbol{\varepsilon}^2] \le 1$. Hence, by Cauchy-Schwartz,

$$\mathscr{R}_n(\mathbb{V}) = \mathbb{E}[\sup_{v\in\mathbb{V}}\frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i v_i] \le \sup_{v\in\mathbb{V}}\|v\|_{2,n}\cdot\sqrt{\frac{1}{n}\mathbb{E}\sum_{i=1}^n \boldsymbol{\varepsilon}_i^2} \le \mathsf{rad}_2(\mathbb{V}).$$

The second bound is a consequence of standard sub-Gaussian maximal inequality (see, e.g. , Theorem 2.5 in Lugosi) and the fact that $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i v_i$ is $\frac{1}{n}\|v\|_{2,n}^2$-subGaussian (e.g., the discussion in Boucheron et al. (2013, Chapter 2.3)). □

We now turn to the proof of Proposition 4.4. We first state two useful lemmas. The first is a direct consequence of Hölder's inequality and the fact that $\|\cdot\|_{p,n} \le \|\cdot\|_{p',n}$ for $p' \ge p$.

Lemma B.6 (Variant of Hölder's inequality)**.** *For any $p, q \ge 2$ satisfying $1/p + 1/q \le 1/2$,*

$$\forall v, u \in \mathbb{R}^n, \quad \|v \odot u\|_{2,n} \le \|v\|_{p,n}\cdot\|u\|_{q,n} \tag{B.10}$$

[3] Recall a variable $\boldsymbol{\varepsilon}$ is 1-subGaussian if, for all $\lambda \ge 0$, $\log\mathbb{E}[\exp(\lambda\boldsymbol{\varepsilon})] \le \lambda^2/2$.

The second bounds the Rademacher complexity of Hadamard products in terms of a finite cover. It is stated with a factor of 2 for convenience when applied below.

Lemma B.7. *Let $\mathbb{V}, \mathbb{U} \subset \mathbb{R}^n$, $p, q \geq 2$ satisfy $1/p + 1/q \leq 1/2$, $\delta_1, \delta_2 \geq 0$, and let $\mathbb{V}'$ be a $2\delta_1$-net of $\mathbb{V}$ in $\|\cdot\|_{p,n}$ and $\mathbb{U}'$ an $2\delta_2$-net of $\mathbb{U}$ in $\|\cdot\|_{q,n}$. Then,*

$$\mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) \leq \mathscr{R}_n(\mathbb{V}' \odot \mathbb{U}') + 2(\mathsf{rad}_p(\mathbb{V})\delta_1 + \mathsf{rad}_q(\mathbb{U})\delta_2),$$

The above bound also holds for the $\mathscr{G}_n$, and more generaly, any analogous complexity using suprema over 1-subGaussian random variables.

Proof of Lemma B.7. Observe that, for any $(v, u) \in \mathbb{V} \times \mathbb{U}$, there exists a $(v', u') \in \mathbb{V}' \times \mathbb{U}'$ with $\|v - v'\|_{p,n} \leq \delta_1$ and $\|u - u'\|_{q,n} \leq \delta_2$. Hence, by Eq. (B.10) followed by Eq. (B.12),

$$\begin{aligned}\|v \odot u - (v') \odot (u')\|_{2,n} &\leq \|v' \odot (v - v')\|_{2,n} + \|v \odot (u - u)'\|_{2,n} \\ &\leq 2(\mathsf{rad}_q(\mathbb{U})\delta_1 + \mathsf{rad}_p(\mathbb{V})\delta_2)\end{aligned}$$

Hence,

$$\begin{aligned}\mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) &= \mathbb{E} \sup_{(v,u)\in\mathbb{V}\times\mathbb{U}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i u_i v_i \\ &\leq \mathbb{E} \sup_{(v,u)\in\mathbb{V}_{j_1}\times\mathbb{U}_{k_1}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i u_i v_i - \mathbb{E} \sup_{(v,u)\in\mathbb{V}_1\times\mathbb{U}_{j_1}} \inf_{(v',u')\in\mathbb{V}\times\mathbb{U}} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i (u_i v_i - u_i' v_i') \\ &\leq \mathscr{R}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) - \mathbb{E} \sup_{w:\|w\|_{2,n}\leq 2(\mathsf{rad}_q(\mathbb{U})\delta_1 + \mathsf{rad}_p(\mathbb{V})\delta_2)} \frac{1}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i w_i \\ &\leq \mathscr{R}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) + 2(\mathsf{rad}_q(\mathbb{U})\delta_1 + \mathsf{rad}_p(\mathbb{V})\delta_2), \qquad \text{(Lemma B.5)}\end{aligned}$$

as needed. □

We now turn to the proof of the main result of this section.

Proof of Proposition 4.4. Recall that $\mathsf{rad}_p(\mathbb{V})$ and $\mathsf{rad}_q(\mathbb{U})$ denote the radii of $\mathbb{V}$ and $\mathbb{U}$ in the $\|\cdot\|_{p,n}$ and $\|\cdot\|_{q,n}$ norms, respectively, assuming $\mathbf{0} \in \mathbb{U} \cap \mathbb{V}$. The only properties of Rademacher variables we use are those assumed by Lemma B.5, i.e. 1-subGaussianity, so our bound holds for Gaussian complexity and other subGaussian ensembles.

We begin with the classical construction of Dudley's integral. Fix $\delta_1 \leq \mathsf{rad}_p(\mathbb{V}), \delta_2 \leq \mathsf{rad}_q(\mathbb{U})$

$$j_1 := \sup\{j : 2^{-j}\mathsf{rad}_p(\mathbb{V}) \geq \delta_1\}, \quad k_1 := \sup\{k : 2^{-}k\mathsf{rad}_q(\mathbb{U}) \geq \delta_2\}$$

For each $j \in [j_1] \cup \{0\}$, let $\mathbb{V}_j$ denote a minimal $2^{-j}\mathsf{rad}_p(\mathbb{V})$ covering of $\mathbb{V}$ in $\|\cdot\|_{p,n}$. Note that since $\mathbf{0} \in \mathbb{V}$, we can take $\mathbb{V}_0 = \{\mathbf{0}\}$, so $|\mathbb{V}_0| = 0$. Define $\pi_{j_1}(v; \mathbb{V}) = v$ for $v \in \mathbb{V}_{j_1}$, and recursively set $\pi_{j-1}(v; \mathbb{V}) \in \arg\min_{v'\in\mathbb{V}_{j-1}} \|v' - \pi_j(v; \mathbb{V})\|_{p,n}$. Set $\Delta_0(v; \mathbb{V}) = \pi_0(v; \mathbb{V})$, and for $j \geq 1$, set $\Delta_j(v; \mathbb{V}) := \pi_j(v; \mathbb{V}) - \pi_{j-1}(v; \mathbb{V})$ for $j \in [j_1]$. Repeat the construction to construct $\mathbb{V}_k$, projection $\pi_k(\mathbb{U})$, and remainders $\Delta_k(u; \mathbb{U})$ analogously, but replacing $\|\cdot\|_{p,n}$ the its conjugate $\|\cdot\|_{q,n}$. The for all $u \in \mathbb{U}_{j_1}, v \in \mathbb{U}_{k_1}$.

$$v = \sum_{j=0}^{j_1} \Delta_j(v; \mathbb{V}), \quad u = \sum_{k=0}^{k_1} \Delta_k(u; \mathbb{U}). \tag{B.11}$$

Lastly, as a shorthand, set

$$\varepsilon_j(\mathbb{V}) := \mathsf{rad}_p(\mathbb{V})2^{-j}, \quad \varepsilon_k(\mathbb{U}) = 2^{-k}\mathsf{rad}_q(\mathbb{U}),$$

noting that

$$\delta_1 \leq \varepsilon_{j_1}(\mathbb{V}) \leq 2\delta_1, \quad \delta_2 \leq \varepsilon_{k_1}(\mathbb{U}) \leq 2\delta_2 \tag{B.12}$$

By Lemma B.7, it suffices to bound $\hat{\mathscr{R}}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1})$. For, $(u, v) \in \mathbb{V}_{j_1} \times \mathbb{U}_{k_1}$,

$$\begin{aligned}
\hat{\mathscr{R}}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) &= \sup_{(u,v)\in\mathbb{V}_{j_1}\times\mathbb{U}_{k_1}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i u \odot v \\
&= \sup_{(u,v)\in\mathbb{V}_{j_1}\times\mathbb{U}_{k_1}} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{j_1}\sum_{k=0}^{k_1} \boldsymbol{\varepsilon}_i \Delta_j(v;\mathbb{U}) \odot \Delta_k(u;\mathbb{U}) \\
&\le \sum_{j=0}^{j_1}\sum_{k=0}^{k_1} \sup_{(u,v)\in\mathbb{V}_{j_1}\times\mathbb{U}_{k_1}} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i \Delta_j(v;\mathbb{U}) \odot \Delta_k(u;\mathbb{U}) \\
&\le \sum_{j=0}^{j_1}\sum_{k=1}^{k_0} \hat{\mathscr{R}}_n(\mathbb{W}_{j,k})
\end{aligned}$$

where we define $\mathbb{W}_{j,k} := \{\Delta_j(v;\mathbb{U}) \odot \Delta_k(u;\mathbb{U}) : v \in \mathbb{V}_{j_1}, u \in \mathbb{U}_{k_1}\}$. Taking expectations yields

$$\mathscr{R}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) \le \sum_{j=0}^{j_1}\sum_{k=0}^{k_1} \mathscr{R}_n(\mathbb{W}_{j,k}) \tag{B.13}$$

From our construction, we can bound

$$\begin{aligned}
\log|\mathbb{W}_{j,k}| \le \log(|\mathbb{V}_j||\mathbb{V}_{j-1}||\mathbb{U}_k||\mathbb{U}_{k-1}|) \le \log(|\mathbb{V}_j|^2|\mathbb{U}_k|^2) &\le \log(2|\mathbb{V}_j|^2|\mathbb{U}_k|^2) \\
&= 2(\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V})) + \mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U})))
\end{aligned}$$

where we use $1 \le |\mathbb{V}_{j-1}| \le |\mathbb{V}_j| = \mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V}))$, and similarly for the sets $\mathbb{U}_k$. Moreover, Eq. (B.10)

$$\begin{aligned}
\mathsf{rad}_2(\mathbb{W}_{j,k}) &= \sup\{\|\Delta_j(v;\mathbb{U}) \odot \Delta_k(u;\mathbb{U})\|_{2,n} : v \in \mathbb{V}_{j_1}, u \in \mathbb{U}_{k_1}\} \\
&\le \sup_{v\in\mathbb{V}_j} \|\Delta_j(v;\mathbb{U})\|_{p,n} \cdot \sup_{u\in\mathbb{U}_k} \|\Delta_k(u;\mathbb{U})\|_{q,n} \\
&\le \varepsilon_j(\mathbb{V})\varepsilon_k(\mathbb{U})
\end{aligned}$$

Thus, Lemma B.5 yields

$$\begin{aligned}
\mathscr{R}_n(\mathbb{W}_{j,k}) &\le \varepsilon_j(\mathbb{V})\varepsilon_k(\mathbb{U})\sqrt{\frac{2}{n}} \cdot \sqrt{(2(\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V})) + \mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U})))} \\
&\le \frac{2}{\sqrt{n}}\left(\varepsilon_j(\mathbb{V})\varepsilon_k(\mathbb{U})\sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V})} + \varepsilon_j(\mathbb{V})\varepsilon_k(\mathbb{U})\sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U}))}\right) \\
&\le \frac{2}{\sqrt{n}}\left(\mathsf{rad}_q(\mathbb{U})2^{-k}\varepsilon_j(\mathbb{V})\sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V})} + \mathsf{rad}_p(\mathbb{V})2^{-j}\varepsilon_k(\mathbb{U})\sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U}))}\right)
\end{aligned}$$

Hence, Eq. (B.13) and evaluating convergent sums yields

$$\begin{aligned}
\mathscr{R}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) &\le \frac{2}{\sqrt{n}}\sum_{j=0}^{j_1}\sum_{k=0}^{k_1}\left(\mathsf{rad}_q(\mathbb{U})2^{-k}\varepsilon_j(\mathbb{V})\sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V})} + \mathsf{rad}_p(\mathbb{V})2^{-j}\varepsilon_k(\mathbb{U})\sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U}))}\right) \\
&\le \frac{4\mathsf{rad}_q(\mathbb{U})}{\sqrt{n}}\sum_{j=1}^{j_1}\varepsilon_j(\mathbb{V})\sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V}))} + \frac{4\mathsf{rad}_p(\mathbb{V})}{\sqrt{n}}\sum_{k=1}^{k_1}\varepsilon_k(\mathbb{U})\sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon_k(\mathbb{U}))}
\end{aligned}$$

where in the second-to-last line, we use that we have $\mathscr{M}_p(\mathbb{V};\varepsilon_0(\mathbb{U})) = \log|\mathbb{V}_0| = 0$.

To simplify, we invoke the following claim.

Claim B.8 (Sum-to-Integral Coversion)**.** *Let ϕ be a non-increasing function, and let $\varepsilon_j = 2^{-j}R$ for some $R > 0$. Then, for $j_a \le j_b$,*

$$\sum_{j=j_a}^{j_b} \varepsilon_j \phi(\varepsilon_j) \le 2 \int_{\varepsilon_{j_b+1}}^{\varepsilon_{j_a}} \phi(\varepsilon)\mathrm{d}\varepsilon = \int_{\varepsilon_{j_b}}^{2\varepsilon_{j_a}} \phi(\varepsilon/2)\mathrm{d}\varepsilon.$$

Proof. The first inequality follows since ϕ is non-increasing, and the second line uses a change of variables

$$\begin{aligned}\sum_{j=j_a}^{j_b} \varepsilon_j \phi(\varepsilon_j) &\le \sum_{j=j_a}^{j_b} \frac{\varepsilon_j}{\varepsilon_j - \varepsilon_{j+1}} \int_{\varepsilon_{j+1}}^{\varepsilon_j} \phi(\varepsilon)\mathrm{d}\varepsilon = 2\sum_{j=j_a}^{j_b} \int_{\varepsilon_{j+1}}^{\varepsilon_j} \phi(\varepsilon)\mathrm{d}\varepsilon \le 2\int_{\varepsilon_{j_b+1}}^{\varepsilon_{j_a}} \phi(\varepsilon)\mathrm{d}\varepsilon \\ &= \int_{2\varepsilon_{j_b+1}}^{2\varepsilon_{j_a}} \phi(\varepsilon/2)\mathrm{d}\varepsilon = \int_{\varepsilon_{j_b}}^{2\varepsilon_{j_a}} \phi(\varepsilon/2)\mathrm{d}\varepsilon.\end{aligned}$$

□

In particular, since metric entropies are non-increasing in their scale factors,

$$\begin{aligned}\sum_{j=1}^{j_1} \varepsilon_j(\mathbb{V})\sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon_j(\mathbb{V}))} &\le \sum_{j=1}^{j_1} \frac{\varepsilon_j(\mathbb{V})}{\varepsilon_j(\mathbb{V}) - \varepsilon_{j+1}(\mathbb{V})} \int_{\varepsilon_{j+1}(\mathbb{V})}^{\varepsilon_j(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon)}\mathrm{d}\varepsilon \\ &= 2\sum_{j=1}^{j_1} \int_{\varepsilon_{j+1}(\mathbb{V})}^{\varepsilon_j(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon)}\mathrm{d}\varepsilon = 2\int_{\varepsilon_{j_1+1}(\mathbb{V})}^{\varepsilon_1(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon)}\mathrm{d}\varepsilon \\ &= 2\int_{\varepsilon_{j_1}(\mathbb{V})/2}^{\mathsf{rad}_p(\mathbb{V})/2} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon)}\mathrm{d}\varepsilon \\ &= \int_{\varepsilon_{j_1}(\mathbb{V})}^{\mathsf{rad}_p(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon/2)}\mathrm{d}\varepsilon \le \int_{\delta_1}^{\mathsf{rad}_p(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon/2)}\mathrm{d}\varepsilon,\end{aligned}$$

where the last inequality uses Eq. (B.12).

Invoking a similar bound for the analogus $\mathbb{U}$-term, we conclude

$$\mathscr{R}_n(\mathbb{V}_{j_1} \odot \mathbb{U}_{k_1}) \le \frac{4\mathsf{rad}_q(\mathbb{U})}{\sqrt{n}} \int_{\delta_1}^{\mathsf{rad}_p(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon/2)}\mathrm{d}\varepsilon + \frac{4\mathsf{rad}_p(\mathbb{V})}{\sqrt{n}} \int_{\delta_2}^{\mathsf{rad}_q(\mathbb{U})} \sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon/2)}\mathrm{d}\varepsilon$$

Combining with Lemma B.7 and taking the infinum over valid δ_1, δ_2,

$$\begin{aligned}\mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) \le \mathsf{rad}_q(\mathbb{U}) &\underbrace{\inf_{\delta_1 \le \mathsf{rad}_p(\mathbb{V})} \left(2\delta_1 + \tfrac{4}{\sqrt{n}} \int_{\delta_1}^{\mathsf{rad}_p(\mathbb{V})} \sqrt{\mathscr{M}_p(\mathbb{V};\varepsilon/2)}\mathrm{d}\varepsilon\right)}_{\mathscr{D}_{n,p}(\mathbb{V})} \\ + \mathsf{rad}_p(\mathbb{V}) &\underbrace{\inf_{\delta_2 \le \mathsf{rad}_q(\mathbb{U})} \left(2\delta_2 + \tfrac{4}{\sqrt{n}} \int_{\delta_2}^{\mathsf{rad}_q(\mathbb{U})} \sqrt{\mathscr{M}_q(\mathbb{U};\varepsilon/2)}\mathrm{d}\varepsilon\right)}_{\mathscr{D}_{n,p}(\mathbb{U})}\end{aligned}$$

□

B.7. Derivation of Lemma 4.1 from Proposition 4.4

We consider the Rademacher complexity, as Proposition 4.4 guarantees the same holds of the Gaussian complexity. Let $\mathbb{U} = \{w \mapsto (1, 1, \ldots, 1) \in \mathbb{R}^n\}$. Applying Proposition 4.4 with the square-Hölder conjugates $p = 2$ and $q = \infty$. As the construction of $\mathbb{U}$ ensures $\mathbb{V} \odot \mathbb{U} = \mathbb{V}$, this yields

$$\mathscr{R}_n(\mathbb{V}) = \mathscr{R}_n(\mathbb{V} \odot \mathbb{U}) \le \mathsf{rad}_\infty(\mathbb{U})\mathscr{D}_{n,2}(\mathbb{V}) + \mathsf{rad}_2(\mathbb{V})\mathscr{D}_{n,\infty}(\mathbb{U}).$$

Notice that $\mathsf{rad}_\infty(\mathbb{U}) = \|(1,1,\dots,1)\|_\infty = 1$. Moreover, the covering number of $\mathbb{U}$ is 1, so its log-covering numbers are zero. Thus, the integral in $\mathscr{D}_{n,\infty}(\mathbb{U})$ vanishes. This concludes the demonstration that

$$\mathscr{R}_n(\mathbb{V}) \le \mathscr{D}_{n,2}(\mathbb{V}) \tag{B.14}$$

As a consequence,

$$\begin{aligned}
\delta_{n,\mathscr{R}}(\mathcal{H}, c) &:= \inf\left\{ r : \mathscr{R}_n(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2c} \right\} \\
&\le \inf\left\{ r : \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2c} \right\} && \text{(Eq. (B.14))} \\
&:= \delta_{n,\mathscr{D}}(\mathcal{H}, c).
\end{aligned}$$

C. Technical Tools

This section enumerates the accompanying technical results applied in the proofs in Appendix B. Whereas Appendix B highlights conceptually novel arguments, this section massages more standard material into the most convenient form for adoption in the prior section.

The results in this section are stated at the following level of generality: Throughout, let $\mathcal{H} : \mathcal{W} \to \mathcal{R}$ denote a class of functions, and P be a measure over $\mathcal{W}$, with Var and $\mathbb{E}$ its corresponding expectation and variance functionals with respect to P. We say $\mathcal{H}$ contains zero if the function $h_0(w) \equiv 0$ lies in $\mathcal{H}$. Many definitions results below involve star-hulls and convex-hulls.

Definition C.1. Let $\mathbb{V} \subset \mathbb{R}^n$. We let $\mathrm{conv}(\mathbb{V})$ denote its convex hull and $\mathrm{star}(\mathbb{V}) := \{t \cdot v : t \in [0,1], v \in \mathbb{V}\}$. Similarly, for a function class $\mathcal{H} : [0,1] \to \mathbb{R}$, we let $\mathrm{star}(\mathcal{H}) := \{t \cdot h, t \in [0,1], h \in \mathcal{H}\}$, *convex hull* as $\mathrm{conv}(\mathcal{H})$ as the minimal convex set containing $\mathcal{H}$.

C.1. Basic Empirical Process Results

Properties of Rademacher and Gaussian complexities. We recall a couple standard facts about the Rademacher complexity. First is that Rademacher complexity is invariant under the convex hull operation, and also under the star-hull operation if the set contains zero.

Lemma C.1 (Convex Hulls)**.** *If $\mathbb{V}' \subset \mathbb{V}$, $\mathscr{R}_n(\mathbb{V}') \le \mathscr{R}_n(\mathbb{V})$. Moreover, $\mathscr{R}_n(\mathbb{V}) = \mathscr{R}_n(\mathrm{conv}(\mathbb{V}))$, and if in addition, $\mathbf{0} \in \mathbb{V}$, $\mathscr{R}_n(\mathbb{V}) = \mathscr{R}_n(\mathrm{star}(\mathbb{V}))$.*

Proof. Recall $\mathscr{R}_n(\mathbb{V}) := \frac{1}{n}\mathbb{E}_{\boldsymbol{\varepsilon}} \sup_{v \in \mathbb{V}} \sum_{i=1}^n \varepsilon_i v_i$. It is then clear that if $\mathbb{V}' \subset \mathbb{V}$, then $\mathscr{R}_n(\mathbb{V}') \le \mathscr{R}_n(\mathbb{V})$. Then is establishes the first point. The second follows because the maximum of a linear function $v \mapsto \sum_{i=1}^n \varepsilon_i v_i$ occurs on the extreme points of $\mathbb{V}$, which are the same as those of $\mathrm{conv}(\mathbb{V})$. Lastly, if $\mathbb{V}$ contains $\mathbf{0}$, $\mathrm{conv}(\mathbb{V}) \supset \mathrm{star}(\mathbb{V}) \supset \mathbb{V}$. □

Next, we state a classical Lipschitz contraction for Rademacher complexity.

Lemma C.2 (Rademacher Contraction, Lemma 29 in (Rakhlin, 2022))**.** *Let ϕ be any L-Lipschitz function, and given $\mathbb{V} \subset \mathbb{R}^n$, let $\phi(\mathbb{V}) := \{\phi(v) : v \in \mathbb{V}\}$. Then, $\mathscr{R}_n(\phi(\mathbb{V})) \le L\mathscr{R}_n(\mathbb{V})$.*

The following lemma is standard (see, e.g. Wainwright (2019, Chapter 14) or, examine the proof of Rakhlin (2022, Lemma 30)).

Lemma C.3. *Let $\mathcal{H}$ be star-shaped. Then $\mathcal{H}(cr) \subset c\mathcal{H}(r)$ for all $c \ge 1$. Hence, for $c \ge 1$, it holds that $\delta_{n,\mathscr{R}}(\mathcal{H}, cB) \le c\delta_{n,\mathscr{R}}(\mathcal{H}, B)$, and similarly $\delta_{n,\mathscr{G}}(\mathcal{H}, c\sigma) \le c\delta_{n,\mathscr{G}}(\mathcal{H}, \sigma)$.*

The following lemma shows a similar property for the Dudley functional, this time without the constraint that $\mathcal{H}$ is star-shaped.

Lemma C.4. *For any class $\mathcal{H}$ and $c \ge 1$, it holds that $\mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le c^{-1}\mathscr{D}_{n,2}(\mathcal{H}[cr, w_{1:n}])$. Thus, for any $c \ge 1$ and $B > 0$, $\delta_{n,\mathscr{D}}(\mathcal{H}, cB) \le c\delta_{n,\mathscr{D}}(\mathcal{H}, B)$.*

Proof. We then have, recalling Definition 4.1 and using $\mathsf{rad}_2(\mathcal{H}[r, w_{1:n}])) = r$ that

$$\begin{aligned}
\mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) &:= \inf_{\delta \le r} \left(2\delta + \frac{4}{\sqrt{n}} \int_{\delta}^{r} \sqrt{\mathscr{M}_2(\mathbb{V}; \varepsilon/2)} \mathrm{d}\varepsilon \right) \\
&= \inf_{\delta \le r} \left(2\frac{c\delta}{c} + \frac{4}{c\sqrt{n}} \int_{c\delta}^{cr} \sqrt{\mathscr{M}_2(\mathbb{V}; \varepsilon/2c)} \mathrm{d}\varepsilon \right) \\
&= \frac{1}{c} \inf_{\delta \le cr} \left(2\delta + \frac{4}{\sqrt{n}} \int_{\delta}^{cr} \sqrt{\mathscr{M}_2(\mathbb{V}; \varepsilon/2c)} \mathrm{d}\varepsilon \right) \\
&\overset{(i)}{\ge} \frac{1}{c} \inf_{\delta \le cr} \left(2\delta + \frac{4}{\sqrt{n}} \int_{\delta}^{cr} \sqrt{\mathscr{M}_2(\mathbb{V}; \varepsilon/2)} \mathrm{d}\varepsilon \right) \\
&= \frac{1}{c} \mathscr{D}_{n,2}(\mathcal{H}[cr, w_{1:n}]),
\end{aligned}$$

where in (i) we use anti-monotonicity of covering numbers $\mathscr{M}_2(\mathbb{V}; \varepsilon/2c) \le \mathscr{M}_2(\mathbb{V}; \varepsilon/2)$ for $c \ge 1$. Recall from Definition 4.1 the definition

$$\begin{aligned}
\delta_{n,\mathscr{D}}(\mathcal{H}, cB) &:= \inf \left\{ r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2cB} \right\} \\
&= \inf \left\{ cr : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[cr, w_{1:n}]) \le \frac{cr^2}{2B} \right\} \\
&= c \inf \left\{ r : c^{-1} \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[cr, w_{1:n}]) \le \frac{r^2}{2B} \right\} \\
&\le c \inf \left\{ r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2B} \right\} \\
&= c\delta_{n,\mathscr{D}}(\mathcal{H}, cB),
\end{aligned}$$

where the inequality above follows from the first part of the lemma. □

Lemma C.5. *Let $\mathcal{H} : \mathcal{W} \to \mathbb{R}$ be an arbitrary class of functions, and let $h_0 : \mathcal{W} \to \mathbb{R}$ be arbitrary. Then, the class $\mathcal{H} + h_0 := \{h + h_0 : h \in \mathcal{H}\}$ satisfies, for all $n \in \mathbb{N}$, $c > 0$, $q \ge 1$, and $w_{1:n} \in \mathcal{W}^n$ the equalities,*

$$\delta_{n,\mathscr{D}}(\mathcal{H} + h_0, c) = \delta_{n,\mathscr{D}}(\mathcal{H}, c), \quad \mathscr{D}_{n,q}((\mathcal{H} + h_0)[w_{1:n}]) = \mathscr{D}_{n,q}(\mathcal{H}[w_{1:n}]).$$

In particular, recalling the notation of Section 4.2, for any $c > 0$, $\delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, c) = \delta_{n,\mathscr{D}}(\mathcal{F}_{\beta}, c)$ and $\sup_{w_{1:n}} \mathscr{D}_{n,\infty}((\mathcal{G} - g_\star)[w_{1:n}])^2 = \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{G}[w_{1:n}])^2$.

Proof. The proof is immediate from the fact that translation by a single element leaves the covering numbers, and hence metric entropies, unchanged. □

C.2. Deviation Inequalities for Empirical Processes

The following is a standard maximal inequality for empirical processes.

Lemma C.6 (Empirical Process Inequality, Theorem 2.3 in (Bousquet, 2002))**.** *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$ and let P be a measure on $\mathcal{W}$ such that $\sup_{h \in \mathcal{H}} |\mathbb{E}_{\mathbf{w} \sim P}[h(\mathbf{w})]| = 0$. Let $\mathbf{Z} := \frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} h(\mathbf{w}_i)$, and let $r^2 := \sup_{h \in \mathcal{H}} \mathbb{E}_{\mathbf{w} \sim P}[h(\mathbf{w})^2]$. Then, for any choice of parameter $\epsilon > 0$.*

$$\mathbb{P}_{\mathbf{w}_{1:n} \sim P} \left[\mathbf{Z} \ge (1 + \epsilon)\mathbb{E}[\mathbf{Z}] \ge r\sqrt{\frac{2\log(1/\delta)}{n}} + \left(\frac{1}{\epsilon} + \frac{1}{3} \right) \frac{B}{n} \log(1/\delta) \right] \le \delta.$$

By examining the proof of Theorem 2.3 in (Bousquet, 2002) from Theorem 2.1 in that same work, one can check that the concusion of Lemma C.6 holds verbtaim in the folllowing more general setup: the class of functions $\tilde{\mathcal{H}} : \mathcal{W} \times [n] \to [-B, B]$ are index-dependent, the process is $\mathbf{Z} := \frac{1}{n} \sup_{\tilde{h} \in \tilde{\mathcal{H}}} \sum_{i=1}^{n} \tilde{h}(\mathbf{w}_i, i)$, and where we define $r^2 := \frac{1}{n} \mathbb{E}[\sum_{i=1}^{n} \tilde{h}(\mathbf{w}_i, i)^2]$ as the average variance. A special case of this generalization applies to Rademacher processes $\mathbf{Z} := \frac{1}{n} \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} \varepsilon_i h(w_i)$, where ε_i plays the roll of the random variable $\mathbf{w}_i$, and where $\tilde{h}(\varepsilon_i, i) = \varepsilon_i h(w_i)$.

Lemma C.7. *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$. Fix any $w_{1:n} \in \mathcal{W}$, and let $\mathbf{Z} := \frac{1}{n}\sup_{h\in\mathcal{H}}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h(w_i)$, and let $r^2 := \sup_{h\in\mathcal{H}}\|h(w_{1:n})\|_{2,n}^2$. Then, for any choice of parameter $\epsilon > 0$.*

$$\mathbb{P}_{\mathbf{w}_{1:n}\sim P}\left[\mathbf{Z} \geq (1+\epsilon)\mathbb{E}[\mathbf{Z}] \geq r\sqrt{\frac{2\log(1/\delta)}{n}} + \left(\frac{1}{\epsilon}+\frac{1}{3}\right)\frac{B}{n}\log(1/\delta)\right] \leq \delta.$$

We shall also need a related lemma that bounds deviations in terms of Rademacher complexity.

Lemma C.8 (Uniform Convergence, Theorem 2.1 in (Bartlett et al., 2005))**.** *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$ be a family of uniformly bounded functions with $|h| \leq B$ and $\sup_{h\in\mathcal{H}}\mathrm{Var}[h^2] \leq r^2$ and let P be a measure over $\mathcal{W}$. Then, with probability at least $1-\delta$, any*

$$\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{n=1}^n \mathbb{E}_{\mathbf{w}\sim P}[h(\mathbf{w})] - h(\mathbf{w}_i) \leq \left(6\mathbb{E}_{\mathbf{w}_{1:n}\sim P}[\mathscr{R}_n(\mathcal{H}[\mathbf{w}_{1:n}])] + r\sqrt{\frac{2\log(1/\delta)}{n}} + \frac{11B\log(1/\delta)}{n}\right)$$

In particular, if $\mathcal{H} = \mathcal{H}(r)$ for some r, then the above holds for $\nu^2 = r^2$.

C.3. Fixed-Design Guarantees

This section concerns various measures of complexity for a function class when its arguments ("design points") $w_{1:n}$ are treated as deterministic. We begin with the following general lemma, which abstracts away the function class $\mathcal{H}[w_{1:n}]$ evaluated on the design points with a set $\mathbb{V} \subset \mathbb{R}^n$. This lemma measure "offset complexities", were a mean zero process involving $v \in \mathbb{V}$ is offset by norms $-\|v\|^2$. This lemma implies important consequences of this lemma for Gaussian and Rademacher compelxities.

Lemma C.9 (Fixed-Design Master Lemma)**.** *Let $\mathbb{V} \subset \mathbb{R}^n$ be a containing $\mathbf{0} \in \mathbb{R}^n$, with $\mathbb{V}[r] := \{v \in \mathbb{V} : \|v\|_{2,n} \leq r\}$, and let $\Phi : \mathcal{U} \to \mathbb{R}$ be an arbitrary function classes (possibly even of cardinality one). Let $\mathbf{u}_1 \ldots, \mathbf{u}_n$ be a random variables taking values in $\mathcal{U}$, and define the processes*

$$\mathbf{Z}(r) := \sup_{v\in\mathbb{V}[r]}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) \qquad \text{(localized maximal process)}$$

$$\mathbf{Y}(\tau) := \sup_{v\in\mathbb{V}}\sup_{\phi\in\Phi}\left\{\frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{1}{2}\|v\|_{2,n}^2\right\}. \qquad \text{(offset maximal process)}$$

Lastly, define a modification of $\mathbf{Y}$ which replaces the offset by $\|v\|_{2,n}^2$ with the offset by r:

$$\tilde{\mathbf{Y}}(r;\tau) := \sup_{v\in\mathbb{V}[r]}\sup_{\phi\in\Phi}\left\{\frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i)\right\} - \frac{r^2}{2} = 2\tau\mathbf{Z}(r) - \frac{r^2}{2}.$$

Then, the following are true.

(a) With probability one,

$$\mathbf{Y}(\tau) = \sup_{r>0}\tilde{\mathbf{Y}}(r;\tau).$$

(b) With probability one,

$$\mathbf{Y}(\tau) \leq \inf\left\{r^2 : \tilde{\mathbf{Y}}(r;\tau) \leq \frac{r^2}{2}\right\}$$

(c) Suppose that, for any choice of $r > 0$, $\mathbf{Z}(r)$ satisfies the following concentration inequality with parameters $c_1 \geq 1$ and $c_2, c_3 > 0$:

$$\mathbb{P}\left[\mathbf{Z}(r) \leq c_1\mathbb{E}[\mathbf{Z}(\tau)] + c_2 r\sqrt{\frac{\log(1/\delta)}{n}} + \frac{c_3\log(1/\delta)}{n}\right] \leq \delta, \qquad \text{(C.1)}$$

Then, for any $\tau \geq 1$ and $\sigma > 0$, the following holds probability $1-\delta$, the following holds

$$\mathbf{Y}(\sigma\tau) \lesssim c_1^2\tau^2\bar{\delta}_n(\sigma)^2 + \frac{(\tau^2\sigma^2c_2^2 + (c_3/c_2)^2)\log(1/\delta)}{n},$$

where

$$\bar{\delta}_n(\sigma) := \inf\left\{r : \mathbb{E}[\mathbf{Z}(r)] \leq \frac{r^2}{2\sigma}\right\}.$$

Thus, by itegrating,

$$\mathbb{E}\mathbf{Y}(\sigma\tau) \lesssim c_1^2\tau^2\bar{\delta}_n(\sigma)^2 + \frac{(\tau^2\sigma^2(c_2^2+1) + c_3)}{n},$$

Proof. We prove the lemma in parts. First, however, we observe that we may assume without loss of generality that $\mathbb{V}$ is star-shaped. Note that $\mathrm{star}(\mathbb{V}[r]) = \mathrm{star}(\mathbb{V})[r]$. Then, by the same logic as in the proof of Lemma C.1, the inclusion $0 \in \mathbb{V}[r]$ implies $\mathrm{conv}(\mathbb{V}[r]) \supset \mathrm{star}(\mathbb{V}[r]) = \mathrm{star}(\mathbb{V})[r] \supset \mathbb{V}[r]$, which establishes

$$\begin{aligned}\mathbf{Z}(r) &:= \sup_{v\in\mathbb{V}[r]}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) = \sup_{v\in\mathrm{conv}(\mathbb{V}[r])}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i)\\ &\geq \sup_{v\in\mathrm{star}(\mathbb{V})[r]}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) \geq \sup_{v\in\mathbb{V}[r]}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) = \mathbf{Z}(r),\end{aligned}$$

so

$$\mathbf{Z}(r) = \sup_{v\in\mathrm{star}(\mathbb{V})[r]}\sup_{\phi\in\Phi}\frac{1}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i).$$

Similarly, one can show that

$$\mathbf{Y}(\tau) := \sup_{v\in\mathrm{star}(\mathbb{V})}\sup_{\phi\in\Phi}\left\{\frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{1}{2}\|v\|_{2,n}^2\right\}, \quad \tilde{\mathbf{Y}}(r;\tau) := \sup_{v\in\mathrm{star}(\mathbb{V})[r]}\sup_{\phi\in\Phi}\left\{\frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i)\right\} - \frac{r^2}{2}.$$

Hence, we can apply the entire lemma to $\mathrm{star}(\mathbb{V})$, and then convert back to $\mathbb{V}$ by the above reduction.

Part (a). As $r^2 \geq \|v\|_{2,n}^2$ for all $v \in \mathbb{V}[r]$, it is immediate that $\mathbf{Y}(\tau) \geq \sup_{r>0}\tilde{\mathbf{Y}}(r;\tau)$. We prove the other direction. Fix an arbitrary $\epsilon > 0$ and suppose that that $v \in \mathbb{V}$ and $\phi \in \Phi$ satisfy

$$\frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{1}{2}\|v\|_{2,n}^2 \geq \mathbf{Y}(\tau) - \epsilon.$$

Letting $r := \|v\|_{2,n}$, it holds that

$$\mathbf{Y}(\tau) - \epsilon \leq \frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{r^2}{2} \leq \tilde{\mathbf{Y}}(r;\tau) \leq \sup_{r>0}\tilde{\mathbf{Y}}(r;\tau).$$

As ϵ was arbitrary, $\mathbf{Y}(\tau) \leq \tilde{\mathbf{Y}}(r;\tau)$.

Part (b). Suppose that r satisfies

$$\tilde{\mathbf{Y}}(r;\tau) \leq 0.$$

Fix $v \in \mathbb{V}$. Since $\mathbb{V}$ is star-shaped, either $v \in \mathbb{V}[r]$, or $\alpha v \in \mathbb{V}[r]$ for $\alpha = r/\|v\|_{2,n} < 1$. In the first case,

$$\begin{aligned}\sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{1}{2}\|v\|_{2,n}^2 &\le \sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) \\ &= \sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - r^2/2 + r^2/2 \\ &\le \underbrace{\tilde{\mathbf{Y}}(r;\tau)}_{\le \frac{r^2}{2}} + \frac{r^2}{2} \le \frac{r^2}{2}.\end{aligned}$$

In the second case, recalling $\alpha = r/\|v\|_{2,n} < 1$,

$$\begin{aligned}\sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n v_i\phi(\mathbf{u}_i) - \frac{1}{2}\|v\|_{2,n}^2 &= \frac{1}{\alpha}\left(\sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n \alpha v_i\phi(\mathbf{u}_i) - \frac{1}{2\alpha}r^2\right) \\ &\le \frac{1}{\alpha}\left(\sup_{\phi\in\Phi} \frac{2\tau}{n}\sum_{i=1}^n \alpha v_i\phi(\mathbf{u}_i) - \frac{1}{2}r^2\right) + \frac{r^2}{2\alpha}(1-\frac{1}{\alpha}) \\ &\le \frac{r^2}{2\alpha} + \frac{r^2}{2\alpha}(1-\frac{1}{\alpha}) = \frac{r^2}{2}\left(2\alpha^{-1} - \alpha^{-2}\right) \le \frac{r^2}{2}\end{aligned}$$

where the second inequality uses $\max_x(2x - x^2) \le 1$ This concludes the proof of part (b).

Part (c). Applying the AM-GM inequality twice to Eq. (C.1), the following holds with probability $1-\delta$,

$$\begin{aligned}\mathbf{Z}(r) &\le c_1\mathbb{E}[\mathbf{Z}(r)] + \frac{r^2}{4\tau\sigma} + \frac{(2\tau\sigma c_2^2 + c_3)\log(1/\delta)}{n} \\ &= c_1\mathbb{E}[\mathbf{Z}(r)] + \frac{r^2}{4\tau\sigma} + \frac{(2\tau^2\sigma^2 c_2^2 + \tau\sigma c_3)\log(1/\delta)}{n\cdot\tau\sigma} \\ &\le c_1\mathbb{E}[\mathbf{Z}(r)] + \frac{r^2}{4\tau\sigma} + \frac{(3\tau^2\sigma^2 c_2^2 + c_3^2/2c_2^2)\log(1/\delta)}{n\cdot\tau\sigma}.\end{aligned}$$

Consequently, setting $\alpha := 8c_1\tau \ge 1$,

$$\begin{aligned}\tilde{\mathbf{Y}}(r;\sigma\tau) = 2\sigma\tau\mathbf{Z}(r) - \frac{r^2}{2} &\le 2c_1\sigma\tau\mathbb{E}[\mathbf{Z}(r)] - \frac{r^2}{4} + \frac{(3\tau^2\sigma^2 c_2^2 + c_3^2/2c_2^2)\log(1/\delta)}{n} \\ &= \frac{\sigma\alpha}{4}\left(\mathbb{E}[\mathbf{Z}(r)] - \frac{r^2}{\alpha 2\sigma}\right) + \frac{(3\tau^2\sigma^2 c_2^2 + c_3^2/2c_2^2)\log(1/\delta)}{n} - \frac{r^2}{8}\end{aligned}$$

To conclude it suffices to show that the above expression is non-positive for the choice

$$r^2 := \max\left\{\frac{(24\tau^2\sigma^2 c_2^2 + 4(c_3/c_2)^2)\log(1/\delta)}{n}, \alpha^2(\bar{\delta}_n(\sigma)^2 + \epsilon)\right\}.$$

Note that this choice makes the second term in the previous display vanishes, so

$$\begin{aligned}\tilde{\mathbf{Y}}(r;\sigma\tau) &\le \frac{\sigma\alpha}{4}\left(\mathbb{E}[\mathbf{Z}(r)] - \frac{r^2}{2\alpha\sigma}\right) \\ &= \frac{\sigma\alpha}{4}\left(\mathbb{E}[\mathbf{Z}(\alpha\bar{\delta}_n(\sigma)^2)] - \frac{\alpha\bar{\delta}_n(\sigma)^2}{2\sigma}\right) \\ &\overset{(i)}{\le} \frac{\sigma\alpha}{4}\left(\alpha\mathbb{E}[\mathbf{Z}(\bar{\delta}_n(\mathbb{V},\Phi,\sigma)^2)] - \frac{\alpha\bar{\delta}_n(\sigma)^2}{2\sigma}\right) \\ &= \frac{\sigma\alpha^2}{4}\left(\mathbb{E}[\mathbf{Z}(\bar{\delta}_n(\sigma)^2)] - \frac{\bar{\delta}_n(\sigma)^2}{2\sigma}\right) \le 0 \le \frac{r^2}{2}\end{aligned}$$

where (i) uses that $\mathbb{V}$ is star-shaped, so $\mathbb{E}[\mathbf{Z}(\mu r)] \le \mu\mathbb{E}[\mathbf{Z}(r)]$ for any $\mu \ge 1$. The proof now follows by subsituting in $\alpha = 8c_1\tau$. □

Consequences of the Master Lemma. Our first consequence is for Gaussian complexities.

Lemma C.10 (Offset Gaussian Complexity Bound)**.** *Let $\mathcal{H} : \mathcal{W} \to \mathbb{R}$ be a function class containing the zero function. Fix any $\delta \in (0,1)$ and $\sigma > 0$ and $\tau \geq 1$. Then, there exists a constant $c > 0$ such that*

$$\sup_{w_{1:n}} \mathbb{P}_{\boldsymbol{\xi}_{1:n}}\left[\sup_{h\in\mathcal{H}} \frac{2\sigma\tau}{n}\sum_{i=1}^{n}\boldsymbol{\xi}_i h(w_i) - \frac{1}{2n}\|h(w_{1:n})\|_{2,n}^2 > c\tau^2\left(\frac{\sigma^2\log(1/\delta)}{n} + \delta_{n,\mathscr{G}}(\mathcal{H},\sigma)^2\right)\right] \leq \delta,$$

where above $\boldsymbol{\xi}_{1:n}$ are i.i.d. standard Normal.

Proof. Recall the set $\mathcal{H}[r, w_{1:n}] = \{h \in \mathcal{H} : \|h\|_{2,n}^2 \leq r^2\}$. Then, the random variable

$$\mathbf{Z}(r) := \frac{1}{n}\sup_{h\in\mathcal{H}[r;w_{1:n}]}\sum_{i=1}^{n}\boldsymbol{\xi}_i h(w_i)$$

satisfies, by Gaussian-Lipschitz concentration (e.g. Boucheron et al. (2013, Chapter 2.3) or Wainwright (2019, Chapter 3)),

$$\mathbb{P}[\mathbf{Z}(r_0) \geq \mathbb{E}_{\boldsymbol{\xi}}\mathbf{Z}(r_0) + r_0\sqrt{2\log(1/\delta)/n}] \leq \delta. \tag{C.2}$$

The bound now follows from Lemma C.9, where $\mathbf{u}_i = \mathbf{x}_i$ and Φ is a singleton consisting of the identity function. □

We establish a similar guarantee for Rademacher variables.

Lemma C.11 (Offset Rademacher Complexity Bound)**.** *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$ be a function class containing zero, and let $\boldsymbol{\varepsilon}_{1:n}$ be i.i.d. Rademacher random variables. Then, for any $\delta \in (0, 1/2)$, $\sigma > 0$ and $\tau \geq 1$, the following holds with probability $1-\delta$*

$$\sup_{h\in\mathcal{H}} \frac{2\sigma\tau}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i h(w_i) - \frac{1}{2n}\sum_{i=1}^{n}h(w_i)^2 \lesssim \frac{(\tau^2\sigma^2 + B^2)\log(1/\delta)}{n} + \tau^2\delta_{n,\mathscr{R}}(\mathcal{H},\sigma)^2$$

In particular, by integrating,

$$\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\sup_{h\in\mathcal{H}} \frac{2\sigma\tau}{n}\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i h(w_i) - \frac{1}{2n}\sum_{i=1}^{n}h(w_i)^2\right] \lesssim \frac{\tau^2\sigma^2 + B^2}{n} + \tau^2\delta_{n,\mathscr{R}}(\mathcal{H},\sigma)^2$$

Proof. Recall the set $\mathcal{H}[r, w_{1:n}] = \{h \in \mathcal{H} : \|h\|_{2,n}^2 \leq r^2\}$ and let

$$\mathbf{Z}(r) := \frac{1}{n}\sup_{h\in\mathcal{H}[r;w_{1:n}]}\sum_{i=1}^{n}\boldsymbol{\xi}_i h(w_i)$$

By Lemma C.7, for some universal constant $c' > 0$,

$$\mathbb{P}\left[\tilde{\mathbf{Z}}(r_0) > c'\left(\mathbb{E}[\tilde{\mathbf{Z}}(r_0)] + r_0\sqrt{\log(1/\delta)/n} + \frac{B\log(1/\delta)}{n}\right)\right] \leq \delta.$$

The bound now follows from Lemma C.9. □

C.4. Random-Design Complexities

The following is an analogue of Lemma C.9 for random design. It's proof is nearly identical, with the key difference between that localization occurs based on the empirical $\mathcal{L}_2$-norm $\mathbb{E}[h(\mathbf{w})]^2$ and not $\|h(w_{1:n})\|_{2,n}^2$. [4]

[4] This remark is under the identification $\mathbb{V} := \mathcal{H}[w_{1:n}]$. We further not that Lemma C.12 implies Lemma C.9 by choosing the measure P to be a dirac-delta. However, to avoid confusion of the subtle differences in localization, we state these two lemmas separately.

Lemma C.12 (Random-Design Master Lemma)**.** *Let P be a measure over random variables $(\mathbf{u}, \mathbf{w})$, and let $\Phi : \mathcal{U} \to \mathbb{R}$ and $\mathcal{H} : \mathcal{W} \to \mathbb{R}$ be function classes, and recall $\mathcal{H}(r) := \{h \in \mathcal{H} : \mathbb{E}_{\mathbf{w}\sim P} h(\mathbf{w})^2 \le r^2\}$. For $(\mathbf{u}_1, \mathbf{w}_1), \dots, (\mathbf{u}_n, \mathbf{w}_n) \overset{\text{i.i.d}}{\sim} P$, define the processes*

$$\mathbf{Z}(r) := \sup_{h\in\mathcal{H}(r)} \sup_{\phi\in\Phi} \frac{1}{n}\sum_{i=1}^n h(\mathbf{w}_i)\phi(\mathbf{u}_i)$$

$$\mathbf{Y}(\tau) := \sup_{h\in\mathcal{H}} \sup_{\phi\in\Phi} \frac{1}{n}\sum_{i=1}^n h(\mathbf{w}_i)\phi(\mathbf{u}_i) - \mathbb{E}[\mathbf{h}(\mathbf{w})^2]$$

$$\tilde{\mathbf{Y}}(r;\tau) := 2\tau\mathbf{Z}(r) - \frac{r^2}{2}.$$

Then, the conclusions of the fixed-design master lemma Lemma C.9 hold verbatim with the above definitions.

Next, we establish two lemmas which give control on the complexities of relevant random-design (i.e. $\mathbf{w}_{1:n} \sim P$)quantities involving quadratic terms such as $\mathbb{E}[h(\mathbf{w})^2]$.

Lemma C.13 (Quadratic Loss Symmetrization)**.** *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$ be a function class containing zero, let P be a distribution over $\mathcal{W}$, and let $\eta > 0$ be arbitrary. Consider the (very similar) terms*

$$T_1(\eta) := \mathbb{E}_{\mathbf{w}_{1:n}\overset{\text{i.i.d}}{\sim}P}\left[\sup_{h\in\mathcal{H}}\left\{\|h(\mathbf{w}_{1:n})\|_{2,n}^2 - (1+\eta)\mathbb{E}_{\mathbf{w}'\sim P}[h(\mathbf{w}')^2]\right\}\right]$$

$$T_2(\eta) := \mathbb{E}_{\mathbf{w}_{1:n}\overset{\text{i.i.d}}{\sim}P}\left[\sup_{h\in\mathcal{H}}\left\{\mathbb{E}_{\mathbf{w}'\sim P}[h(\mathbf{w}')^2] - (1+\eta)\|h(\mathbf{w}_{1:n})\|_{2,n}^2\right\}\right]$$

$$T_3(\eta) := \mathbb{E}_{\mathbf{w}_{1:n},\mathbf{w}'_{1:n}\overset{\text{i.i.d}}{\sim}P}\left[\sup_{h\in\mathcal{H}}\left\{\|h(\mathbf{w}_{1:n})\|_{2,n}^2 - (1+\eta)\|h(\mathbf{w}'_{1:n})\|_{2,n}^2\right\}\right],$$

as well as the term

$$T_4(\eta) := \mathbb{E}_{\mathbf{w}_{1:n}}\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{\boldsymbol{\varepsilon}_{1:n}}\left[\frac{B(1+\eta)}{n}\sum_{i=1}^n \boldsymbol{\varepsilon}_i h(\mathbf{w}_i)\right] - \frac{1}{2}\mathbb{E}[\|h(\mathbf{w}_{1:n})\|_{2,n}^2]\right]$$

Then,

$$\max\{T_1(\eta), T_2(\eta), T_3(\eta), T_4(\eta)\} \lesssim \frac{\eta(1+\eta^{-1})^{-2}}{n}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H}, B)^2\right).$$

Proof. By Liang et al. (2015, Lemma 14) (modifying the constant of $4B$ to $2B$ to account for the fact that we consider the uncentered $h \in \mathcal{H}$, and not centered $h - h_\star \in \mathcal{H}$, and reparameterizing $\eta \leftarrow \eta/2$), it holds that

$$T_2(\eta) \le \frac{\eta}{2n}\mathbb{E}_{\mathbf{w}_{1:n}\sim P}\left[\sum_{i=1}^n \frac{2B(2+\eta)}{\eta}\boldsymbol{\varepsilon}_i h(\mathbf{w}_i) - h(\mathbf{w}_i)^2\right]$$

The same argument can be modified to show that $T_1(\eta)$ satisfies the same upper bound, as $T_1(\eta)$ satisfies the same intermediate inequality obtained via Jensen's inequality (the third line of the proof in Liang et al. (2015, Lemma 14)), and the same argument extends to $T_3(\eta)$ because this expresssion is precisely the consequence of applying Jensen's inequality. Thus,

$$\begin{aligned}
\max\{T_1(\eta), T_2(\eta), T_3(\eta)\} &\le \frac{\eta}{2n}\mathbb{E}_{\mathbf{w}_{1:n}\sim P}\left[\sum_{i=1}^n \frac{2B(1+2\eta)}{\eta}\boldsymbol{\varepsilon}_i h(\mathbf{w}_i) - h(\mathbf{w}_i)^2\right] \\
&\le \frac{2\eta}{n}\mathbb{E}_{\mathbf{w}_{1:n}\sim P}\left[\sum_{i=1}^n 4B(2+\eta^{-1})\boldsymbol{\varepsilon}_i h(\mathbf{w}_i) - \frac{1}{2}h(\mathbf{w}_i)^2\right] \\
&\overset{(i)}{\lesssim} \frac{\eta(2+\eta^{-1})^2}{n}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H}, B)^2\right), \\
&\lesssim \frac{\eta(1+\eta^{-1})^2}{n}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H}, B)^2\right),
\end{aligned}$$

where the inequality (i) is by Lemma C.11 with $\sigma = B$, and $\tau = 2(1+\eta^{-1})$. The bound on $T_4(\eta)$ follows from a similar application of Lemma C.11. □

Lemma C.14 (Quadratic Lower Bound). *Let $\mathcal{H} : \mathcal{W} \to [-B, B]$ be a function class containing zero, and let P be a measure over $\mathcal{W}$. Then, there is a universal constant $c > 0$ such that for any $\delta > 0$, it holds that*

$$P\left[\sup_{h\in\mathcal{H}} \|h\|^2_{\mathcal{L}_2(P)} - 2\|h(\mathbf{w}_{1:n})\|^2_{2,n} \le c r_{\mathrm{quad}}(\mathcal{H},\delta)^2\right] \le 1-\delta,$$

where

$$r_{\mathrm{quad}}(\mathcal{H},\delta)^2 := \left(\delta_{n,\mathscr{R}}(\mathcal{H},B)^2 + \frac{B^2\log(1/\delta)}{n}\right).$$

Proof. In view of Lemma C.1, the fact that $\mathcal{H}$ contains zero means we may assume without loss of generality that $\mathcal{H}$ is star-shaped (indeed, apply the lemma to $\mathrm{star}(\mathcal{H})$, and note that $\delta_{n,\mathscr{R}}(\mathcal{H},B) = \delta_{n,\mathscr{R}}(\star(\mathcal{H}),B)$). Introduce the class of function $\tilde{\mathcal{H}}_r := \{\mathbb{E}[h^2] - h^2 : h \in \mathcal{H}(r)\}$ (here, we use subscript r to distinguish from the standard localization notation). Then, $\tilde{\mathcal{H}}_r : \mathcal{W} \to [-B^2, B^2]$, and $\mathbb{E}[\tilde{h}(\mathbf{w})^2] \le B^2 r^2$ for $\tilde{h} \in \tilde{\mathcal{H}}_r$. Note that

$$\sup_{h\in\tilde{\mathcal{H}}_r} \frac{1}{n}\sum_{i=1}^n \tilde{h}(\mathbf{w}_i) = \sup_{h\in\mathcal{H}(r)} \|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n}$$

Hence, by Lemma C.6 and AM-GM, the following holds with probability $1-\delta$ and for a universal constant $c > 0$ and any $\tau \ge 0$:

$$\begin{aligned}
&\sup_{h\in\mathcal{H}(r)} \|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n} \\
&\le c\left(\mathbb{E}\left[\sup_{h\in\tilde{\mathcal{H}}_r} \|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n}\right] + \tau r^2 + \frac{(\tau^{-1}+1)B^2\log(1/\delta)}{n}\right) \\
&\le c\left(\mathbb{E}\left[\sup_{h\in\mathcal{H}(r)} (1-\tau)\|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n}\right] + 2\tau r^2 + \frac{(\tau^{-1}+1)B^2\log(1/\delta)}{n}\right),
\end{aligned}$$

where the second inequality uses $\|h\|^2_{\mathcal{L}_2(P)} \le r^2$. Let $\tau \le 1/2$ and let η be such that $(1-\tau) = \frac{1}{1+\eta}$. By Lemma C.13,

$$\begin{aligned}
\mathbb{E}\left[\sup_{h\in\mathcal{H}(r)} (1-\tau)\|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n}\right] &= \frac{1}{1+\eta}\mathbb{E}\left[\sup_{h\in\mathcal{H}(r)} \|h\|^2_{\mathcal{L}_2(P)} - (1+\eta)\|h(\mathbf{w}_{1:n})\|^2_{2,n}\right] \\
&\lesssim \frac{\eta(1+\eta^{-1})^{-2}}{(1+\eta)}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H},B)\right) \\
&= (1-\eta^{-1})\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H},B)^2\right) \\
&\lesssim \frac{1}{\tau}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H},B)^2\right),
\end{aligned}$$

where in the last line, we use that $\eta = 1-(1-\tau)^{-1} \gtrsim \tau$ for $\tau \le 1/2$. In sum, there is a universal constant c' such that, for all $\tau \le 1/2$, the following holds with probability $1-\delta$:

$$\sup_{h\in\mathcal{H}(r)} \|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n} \le c'\left(\frac{1}{\tau}\left(\frac{B^2}{n} + \delta_{n,\mathscr{R}}(\mathcal{H},B)\right)^2 + \tau r^2 + \frac{(\tau^{-1}+1)B^2\log(1/\delta)}{n}\right).$$

By making τ a sufficiently small universal constant, we can ensure that there is a universal constants c'', c''' such that, whenever

$$r^2 = c''\left(\delta_{n,\mathscr{R}}(\mathcal{H},B)^2 + \frac{B^2\log(1/\delta)}{n}\right) \lesssim r_{\mathrm{quad}}(\mathcal{H},\delta).$$

we have that with probability $1-\delta$,

$$\sup_{h\in\mathcal{H}(r)} \|h\|^2_{\mathcal{L}_2(P)} - \|h(\mathbf{w}_{1:n})\|^2_{2,n} \le \frac{r^2}{2}. \tag{C.3}$$

We claim that in fact, with probability $1-\delta$, it holds that the above holds for all $h\in\mathcal{H}$, that is

$$\sup_{h\in\mathcal{H}} \|h\|^2_{\mathcal{L}_2(P)} - 2\|h(\mathbf{w}_{1:n})\|^2_{2,n} \le \frac{r^2}{2}$$

Indeed, it suffices to check Eq. (C.3) implies the inequality for $h\notin\mathcal{H}(r)$. Since $\mathcal{H}$ is star-shaped, there exists some α such that $\alpha h\in\mathcal{H}(r)$ and in fact $\|\alpha h\|^2_{\mathcal{L}_2(P)} = r^2$. Then, on Eq. (C.3)

$$\|\alpha h\|^2_{\mathcal{L}_2(P)} - \|\alpha h(\mathbf{w}_{1:n})\|^2_{2,n} \le \frac{r^2}{2} = \frac{\|\alpha h\|^2_{\mathcal{L}_2(P)}}{2}$$

so by rearranging

$$\|h\|^2_{\mathcal{L}_2(P)} \le 2\|h(\mathbf{w}_{1:n})\|^2_{2,n}.$$

□

D. Rates for finite function classes.

In this section, we establish sharper bounds for finite function classes $\mathcal{H}_1$ and $\mathcal{H}_2$. Because errors for finite function classes already attain the $\mathcal{O}(1/n)$ parametric rate, we require an additional assumption to achieve improvement. Specifically, we need a hypercontractivity condition which states that higher moments of $\mathbb{E}[h(\mathbf{w})^q]$ for $h\in\mathcal{H}$ are controller by lower order moments $\mathbb{E}[h(\mathbf{w}^s)]$ for $s<q$. This is the first notion of hypercontractivity, defined below.

Definition D.1 (Hypercontractivity)**.** We say a class $\mathcal{H}\subset\{\mathcal{W}\to\mathbb{R}\}$ satisfies $(\kappa,\mathbb{P},s,q)$-hypercontractivity if, for all $h\in\mathcal{H}$, $\mathbb{E}[|h(\mathbf{w})|^q]^{1/q}\le\kappa\mathbb{E}[h(\mathbf{w})^s]^{1/s}$.

We achieve even faster rates under a stronger variant of hypercontractivity, defined below.

Definition D.2 (subGaussian Hypercontractivity)**.** We say a class $\mathcal{H}\subset\{\mathcal{W}\to\mathbb{R}\}$ satisfies $(\kappa,\mathbb{P})$-subGaussian hypercontractivity if, for all $h\in\mathcal{H}$, $h(\mathbf{w})-\mathbb{E}[h(\mathbf{w})]$ is $\kappa^2\mathbb{E}[h(\mathbf{w})^2]$ subGaussian. [5]

Under the various hypercontractivity assumptions, we attain the following bound, which is the formal statement of Theorem 2.

Theorem 4. *Let $\mathcal{F}:\mathcal{X}\to[-1,1]$ and $\mathcal{G}:\mathcal{Y}\to[-1,1]$ be finite function classes, and suppose $1\le d_1\le d_2$ satisfy $\log|\mathcal{F}|\le d_1$ and $\log|\mathcal{G}|\le d_2$. Define the class[6] $\tilde{\mathcal{F}}:=\{f-\beta_f-f_\star : f\in\mathcal{F}\}$, and casing on the hypercontractivity assumptions with parameter κ, define*

$$\phi_n(d_1,d_2) := \begin{cases} \left(\frac{d_2}{n}\right)^{\frac{2}{q_2}} + \left(\frac{d_1}{d_2}\right)^{\frac{1}{q_1}} & \text{general } \frac{1}{q_1}+\frac{1}{q_2}=2,\ \tilde{\mathcal{F}} \text{ satisfies } (\kappa,\mathbb{P}_{\mathsf{train}},2,q_1) \text{ hypercontractivity} \\ \left(\frac{d_2}{n}\right)^{\frac{1}{2}} + \left(\frac{d_1}{d_2}\right)^{\frac{1}{4}} & \tilde{\mathcal{F}} \text{ satisfies } (\kappa,\mathbb{P}_{\mathsf{train}},2,4) \text{ hypercontractivity} \\ \frac{d_2}{n}\cdot(d_1+\log n) & \tilde{\mathcal{F}} \text{ satisfies } (\kappa,\mathbb{P}_{\mathsf{train}})\text{-subGaussian hypercontractivity} \end{cases}$$

Then, as long as $\sigma^2\lesssim 1$, for any $\delta\in(e^{-10d_2},e^{-1})$, the following hold simultaneously with probability at least $1-\delta$:

$$\mathcal{R}_{\mathrm{train}}[\hat{g}_n;\hat{f}_n] \le \mathcal{R}_{\mathrm{train}}(\hat{f}_n,\hat{g}_n) \lesssim \frac{d_2}{n}$$

$$\mathcal{R}_{\mathsf{train}}[\hat{f}_n] \lesssim \kappa^2\phi_n(d_1,d_2)\cdot\frac{d_2}{n} + \frac{d_1}{n} + \frac{\log(1/\delta)}{n}$$

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n,\hat{g}_n) \lesssim \nu_{x,y}\cdot\frac{d_1+\log(1/\delta)}{n} + (\nu_y+\nu_{x,y}\cdot\kappa^2\phi_n(d_1,d_2))\cdot\frac{d_2}{n}$$

[5] This is equivalent to $\mathcal{O}(\kappa)$-hypercontractivity in the $\mathcal{L}_2\to\psi_2$ norms, where ψ_2 is the subGaussian (Orlicz) norm (see e.g. Boucheron et al. (2013, Exercise 2.18).)

[6] note that $\mathcal{F}_{\mathtt{cnt}} := \mathrm{star}(\tilde{\mathcal{F}})$

Notice that, as promised by Theorem 2 $\phi_n(d_1, d_2)$ tends to 0 as $n \to \infty$ and as the ratio of the class complexities d_1/d_2 tends to 0, such that with high probability. Moreover, under subGaussian hypercontractivity, $\lim_{n\to\infty} \phi_n(d_1, d_2)$ for any d_1, d_2 fixed.

Remark D.1 (Extension to Parametric Classes)**.** Up to logarithmic factors in n, the above bound can be extended easily extended to infinite-cardinality "parametric" function classes (that is, function classes whose metric entropies scale as logarithmic in the scale ϵ). The guarantee of Theorem 4 also holds under the more general assumption that $\mathcal{F}$ and $\mathcal{G}$ are contained in the respective convex hulls of function classes $\tilde{\mathcal{F}}$ and $\tilde{\mathcal{G}}$, where $\log|\tilde{\mathcal{F}}| \le d_1$ and $\log|\tilde{\mathcal{G}}| \le d_2$. This includes, for example, many natural linear classes.

D.1. Proof of Theorem 4

We start with localization for finite classes.

Lemma D.1 (Localization for Finite Classes)**.** *Let $\mathcal{H}$ be a finite function class uniformly bounded by 1, and let $d = \log|\mathcal{H}|$. Then, for any probability measure $\mathbb{P}$ over $\mathcal{W}$,*

$$\delta^2_{n,\mathscr{R}}(\mathcal{H}, B) \lesssim \frac{B^2 d}{n}, \quad \delta_{n,\mathscr{G}}(\mathcal{H}, \sigma) \lesssim \frac{d\sigma^2}{n}$$

Proof of Lemma D.1. From Lemma C.8 and finiteness of $\mathcal{H}$, we have for any $w_{1:n} \in \mathcal{W}^n$ that

$$\mathscr{R}_n(\mathcal{H}[r; w_{1:n}]) \le \mathsf{rad}_2(\mathcal{H}[r; w_{1:n}]\sqrt{2d/n} \le r\sqrt{2d/n}.$$

It then follows that $\delta^2_{n,\mathscr{R}}(\mathcal{H}, B) = \sup\{r^2 : \mathscr{R}_n(\mathcal{H}[r; w_{1:n}]) \le \frac{r^2}{2B}\} \lesssim \frac{B^2 d}{n}$. The bound on $\delta_{n,\mathscr{G}}$ similarly yields $\delta^2_{n,\mathscr{R}}(\mathcal{H}) = \sup\{r^2 : \mathscr{R}_n(\mathcal{H}[r; w_{1:n}]) \le \frac{r^2}{2\sigma}\} \lesssim \frac{\sigma^2 d}{n}$. □

We continue with a generic bound on the following cross-critical radius. The next proposition is proved in Appendix D.2 below.

Proposition D.2. *For $i \in \{1, 2\}$, let $\mathcal{H}_i \subset \{\mathcal{W} \to [-1, 1]\}$ be finite function classes with $d_i = \log|\mathcal{H}_i|$. Assume for simplicity that $d_1 \le d_2$, and let $\gamma^2 \ge d_2/n$. Finally, let $\mathbb{P}$ be a distribution of $\mathcal{W}$. Define the shorthand*

$$\delta_n(\gamma) := \inf\left\{r : \mathbb{E}_{\mathbf{w}_{1:n}}[\mathscr{R}_n((\mathcal{H}_1(r) \odot \mathcal{H}_2(\gamma))[\mathbf{w}_{1:n}])] \le \frac{r^2}{2}\right\}.$$

Then, it holds that

- *Let $1/q_1 + 1/q_2 = 1/2$ be square Hölder conjugates. If $\mathcal{H}_1$ satisfies $(\kappa, \mathbb{P}, 2, q_1)$ hypercontractivity,*

$$\delta_n(\gamma)^2 \lesssim \kappa^2\gamma^{4/q_2}\frac{d_2}{n} + \gamma^{2/q_2}\sqrt{\frac{d_2}{n}}\left(\frac{d_1}{n}\right)^{1/q_1}.$$

  In particular, if $\gamma^2 \simeq d_2/n$,

$$\delta_n(\gamma)^2 \lesssim \kappa^2 \begin{cases} \left(\frac{d_2}{n}\right)^{1+\frac{2}{q_2}} + \frac{d_2}{n}\cdot\left(\frac{d_1}{d_2}\right)^{\frac{1}{q_1}} & \text{general } \frac{1}{q_1}+\frac{1}{q_2} = 2 \\ \left(\frac{d_2}{n}\right)^{\frac{3}{2}} + \frac{d_2}{n}\cdot\left(\frac{d_1}{d_2}\right)^{\frac{1}{4}} & q_1 = q_2 = 4. \end{cases}$$

- *If $\mathcal{H}_1$ satsfies $(\kappa, \mathbb{P})$-subGaussian hypercontractivity,*

$$\delta_n(\gamma)^2 \lesssim \frac{\kappa^2\gamma^2 d_2}{n}\cdot(d_1 + \log n).$$

  In particular, if $\gamma^2 \simeq d_2/n$, the above scales as $\kappa^2(d_2/n)^2\cdot(d_1 + \log n)$.

Next, we recall standard localization bounds for finite function classes.

Proof of Theorem 4. As $d_2 \geq d_1$, $\log|\mathcal{F}+\mathcal{G}| \leq \log(|\mathcal{F}|+|\mathcal{G}|) \lesssim d_2$. Taking $B=1$, $\sigma^2 \lesssim 1$, and $\delta = e^{-d_2}$, Lemma D.1 and Proposition 4.2 allow us to bound

$$\mathcal{R}_{\mathrm{train}}[\hat{g}_n; \hat{f}_n] \leq \mathcal{R}_{\mathrm{train}}(\hat{f}_n, \hat{g}_n) \lesssim \frac{d_2}{n} \text{ w.p. } 1-e^{-d_2}$$

By the same token, applying Proposition 4.3 and Lemma D.1, and making similar simplifications ($B=1, \sigma^2 \lesssim 1$), the following holds with probabilty $1-\delta$

$$\mathcal{R}_{\mathsf{train}}[\hat{f}_n] \lesssim \bar{\delta}_{n,\mathrm{cross}}(\gamma)^2 + \frac{d_1+\log(1/\delta)}{n}$$

Bounding $\bar{\delta}_{n,\mathrm{cross}}(\gamma)^2$ by Proposition D.2, on the same event we have

$$\mathcal{R}_{\mathsf{train}}[\hat{f}_n] \lesssim \frac{\kappa^2\phi_n(d_1,d_2)d_2}{n} + \frac{d_1+\log(1/\delta)}{n},$$

where we recall $\phi_n(d_1,d_2)$ defined in the theorem statement. When both events hold, Lemma 3.1 entails

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim \nu_{x,y}\left(\frac{\kappa^2\phi_n(d_1,d_2)d_2}{n} + \frac{d_1}{n} + \frac{\log(1/\delta)}{n}\right) + \nu_y\frac{d_2}{n}.$$

□

D.2. Proof of Proposition D.2

Define the radius of a class $\mathcal{H}: \mathcal{W} \to \mathbb{R}$ be a class, and let P be a measure over $\mathcal{W}$. Define

$$\overline{\mathsf{rad}}_q(\mathcal{H}) := \sup_{h\in\mathcal{H}} \mathbb{E}[|h(\mathbf{w})|^q]^{1/q}, \quad \mathsf{rad}_q(\mathcal{H}[w_{1:n}]) := \sup_{h\in\mathcal{H}} \|h[w_{1:n}]\|_{q,n}$$

Part 1. Bounds on the empircal norms. The next lemma bounds the magnitude of the empirical q-norm radius.

Lemma D.3. *Let $\mathcal{H} \subset \{\mathcal{W} \to [-1,1]\}$ be a finite class, and take $\delta \in (0,1)$. With probability at least $1-\delta$,*

$$\mathsf{rad}_q(\mathcal{H}[\mathbf{w}_{1:n}]) \leq 2\overline{\mathsf{rad}}_q(\mathcal{H}) + \left(\frac{\log|\mathcal{H}|/\delta}{n}\right)^{\frac{1}{q}} \leq 1-\delta.$$

In particular, if $\overline{\mathsf{rad}}_2(\mathcal{H}) \leq r$ and $\mathcal{H}$ satisfes $(\kappa, \mathbb{P}, 2, q)$ hypercontractivity, then with probability $1-\delta$

$$\mathsf{rad}_q(\mathcal{H}[\mathbf{w}_{1:n}]) \leq 2\kappa r + \left(\frac{\log 1/\delta}{n}\right)^{\frac{1}{q}}.$$

Proof. We observe that $\mathsf{rad}_q(\mathcal{H}[\mathbf{w}_{1:n}])^q \leq \overline{\mathsf{rad}}_q(\mathcal{H})^q + \mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n |h(\mathbf{w}_i)|^q - \mathbb{E}[|h(\mathbf{w}_i)^q|]$. As $\sup_h |h| \leq 1$, $\sup_{h\in\mathcal{H}} \mathbb{E}[|h(\mathbf{w}_i)|^{2q}] \leq \sup_{h\in\mathcal{H}} \mathbb{E}[|h(\mathbf{w}_i)|^q] = \overline{\mathsf{rad}}_q(\mathcal{H})^q$. By Bernstein's inequality and a union bound, with probability at least $1-\delta$,

$$\begin{aligned}\mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n |h(\mathbf{w}_i)|^q - \mathbb{E}[|h(\mathbf{w}_i)|^q] &\leq \sqrt{\frac{2\overline{\mathsf{rad}}_q(\mathcal{H})^q \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log|\mathcal{H}|/\delta}{3n} \\ &\leq \overline{\mathsf{rad}}_q(\mathcal{H})^q + \frac{\log|\mathcal{H}|/\delta}{n}. \qquad \text{(AM-GM, and } \tfrac{1}{3}+\tfrac{1}{2} \leq 1)\end{aligned}$$

Hence, via the previous two displays, with probability $1-\delta$ it holds that

$$\mathsf{rad}_q(\mathcal{H}[\mathbf{w}_{1:n}])^q \leq 2\overline{\mathsf{rad}}_q(\mathcal{H})^q + \frac{\log|\mathcal{H}|/\delta}{n}.$$

Taking the q-th root and using of $(x+y)^{1/q} \leq y^{1/q} + x^{1/q}$ for $q \geq 1$ and $x, y \geq 0$ concludes the proof. □

When $\mathcal{H}$ satisfies $(\kappa, \mathbb{P})$-subGaussian hypercontractivity, we can improve this bound.

Lemma D.4. *Suppose $\mathcal{H} \subset \{\mathcal{W} \to [-1,1]\}$ be a finite class which satisfies $(\kappa, \mathbb{P})$-subGaussian hypercontractivity. Then,*

$$\mathbb{P}[\mathsf{rad}_\infty(\mathcal{H}[\mathbf{w}_{1:n}]) \le \kappa\sqrt{\log|\mathcal{H}| + \log(n/\delta)} \cdot \overline{\mathsf{rad}}_2(\mathcal{H})] \ge 1-\delta.$$

Proof. By Gaussian concentration and $(\kappa, \mathbb{P})$-suBgaussian hypercontractivity, for any $h \in \mathcal{H}$, $i \in [n]$ and $\delta > 0$, we have

$$\mathbb{P}\left[|h(\mathbf{w})_i| \ge \kappa\overline{\mathsf{rad}}_2(\mathcal{H})\log(1/\delta)\right] \le \mathbb{P}\left[|h(\mathbf{w})_i| \ge \kappa\mathbb{E}[h(\mathbf{w}_i)^2]^{1/2}\log(1/\delta)\right] \le \delta$$

Union bounding over $h \in \mathcal{H}$ and $i \in [n]$ concluds the proof. □

Part 2. Controlling the Rademacher Complexities Next, we turn to bounding the Rademacher complexity in terms of empirical radii.

Lemma D.5. *Let $1/q_1 + 1/q_2 \le 1/2$ be squared Hölder conjugates. Then,*

$$\mathscr{R}_{n,\mathbb{P}}(\mathcal{H}_1(r) \odot \mathcal{H}_2(\gamma)) \le \sqrt{2(d_1+d_2)}\mathbb{E}\left[\mathsf{rad}_{q_1}(\mathcal{H}_{1,r}[\mathbf{w}_{1:n}]) \cdot \mathsf{rad}_{q_2}(\mathcal{H}_{2,\gamma}[\mathbf{w}_{1:n}])\right]$$

Proof. This is a direct consequence of Lemma B.5, and the fact that $\log|\mathcal{H}_1(r) \odot \mathcal{H}_2(\gamma)| \le \log|\mathcal{H}_1| + \log|\mathcal{H}_2| = d_1 + d_2$. □

Part 3. Conclusion the proof. We can now conclude.

Proposition D.2. Let us start with the case that $\mathcal{H}_1$ satisfies $(\kappa, \mathbb{P}, 2, q_1)$ hypercontractivity. Uing boundedness of $|h| \le 1$ and $q_2 \le 1$ we get

$$\overline{\mathsf{rad}}_{q_2}(\mathcal{H}_{2,\gamma})^{q_2} = \sup_{h\in\mathcal{H}_{2,\gamma}} \mathbb{E}[|h(\mathbf{w})|^{q_2}] \le \sup_{h\in\mathcal{H}_{2,\gamma}} \mathbb{E}[|h(\mathbf{w})|^2] \le \gamma^2,$$

so $\overline{\mathsf{rad}}_{q_2}(\mathcal{H}_{2,\gamma}) \le \gamma^{2/q_2}$. Hence, as $d_2/n \le \gamma^2$ by assumption,

$$\overline{\mathsf{rad}}_{q_2}(\mathcal{H}_{2,\gamma}) + (\frac{d_2}{n})^{\frac{1}{q_2}} \le \gamma^{2/q_2} + (d_2/n)^{1/q_2} \le 2\gamma^{2/q_2}. \tag{D.1}$$

Next, by Lemma D.3, hypercontractivity of $\mathcal{H}_1$, and the above bound implies that, for all $\delta > 0$, both

$$\mathsf{rad}_{q_1}(\mathcal{H}_{1,r}[\mathbf{w}_{1:n}]) \le 2\kappa_1 r + \left(\frac{d_1 + \log 1/\delta}{n}\right)^{\frac{1}{q_1}}$$

and

$$\mathsf{rad}_{q_2}(\mathcal{H}_{1,r}[\mathbf{w}_{1:n}]) \le 2\gamma^{2/q_2} + \left(\frac{d_2 + \log 1/\delta}{n}\right)^{\frac{1}{q_2}},$$

hold with probability at least $1-\delta$. Taking the product, integrating the tail over δ, and invoking Eq. (D.1) implies

$$\mathbb{E}\left[\mathsf{rad}_{q_1}(\mathcal{H}_{1,r}[\mathbf{w}_{1:n}]) \cdot \mathsf{rad}_{q_2}(\mathcal{H}_{2,\gamma}[\mathbf{w}_{1:n}])\right] \lesssim \gamma^{2/q_2}\left(r\cdot\kappa + \left(\frac{d_1}{n}\right)^{1/q_1}\right).$$

Note the resulting constant does not depend on q_1 or q_2, as $0 \le 1/q_1, 1/q_2 \le 1$ are both bounded. Consequently, Lemma D.5 and $d_1 \le d_2$ entails

$$\mathscr{R}_{n,\mathbb{P}}(\bar{\mathcal{H}}_1(r) \odot \bar{\mathcal{H}}_2(\gamma)) \lesssim \sqrt{\frac{d_2}{n}}\gamma^{2/q_2}\left(r\cdot\kappa + \left(\frac{d_1}{n}\right)^{1/q_1}\right),$$

Thus,

$$\inf\left\{r^2 : \mathscr{R}_{n,\mathbb{P}}(\bar{\mathcal{H}}_1(r)\odot\bar{\mathcal{H}}_2(\gamma)) \le \frac{r^2}{2}\right\} \lesssim \kappa^2\gamma^{4/q_2}\frac{d_2}{n} + \gamma^{2/q_2}\sqrt{\frac{d_2}{n}}\left(\frac{d_1}{n}\right)^{1/q_1}.$$

Next, let's consider the subGaussian hypercontractive case. Here, we replace Lemma D.3 with Lemma D.4. A similar computation yields

$$\mathbb{E}\left[\mathsf{rad}_\infty(\mathcal{H}_{1,r}[\mathbf{w}_{1:n}])\cdot\mathsf{rad}_2(\mathcal{H}_{2,\gamma}[\mathbf{w}_{1:n}])\right] \lesssim \kappa r\gamma\sqrt{d_1+\log n}.$$

Hence, Lemma D.5 (with $d_1 \le d_2$) gives

$$\mathscr{R}_{n,\mathbb{P}}(\bar{\mathcal{H}}_1(r)\odot\bar{\mathcal{H}}_2(\gamma)) \lesssim \sqrt{\frac{d_2}{n}}\kappa r\gamma\sqrt{d_1+\log n}.$$

We may then conclude

$$\inf\left\{r^2 : \mathscr{R}_{n,\mathbb{P}}(\bar{\mathcal{H}}_1(r)\odot\bar{\mathcal{H}}_2(\gamma)) \le \frac{r^2}{2}\right\} \lesssim \frac{\kappa^2\gamma^2 d_2}{n}\cdot\kappa(d_1+\log n).$$

□

E. Formal Guarantees for Nonparametric Classes (formal statement of Theorem 1)

In this section, give a formal statement of Theorem 1. Recall the definition of the normalized q-norms: For $v = (v_1,\dots,v_n)\in\mathbb{R}^n$ and $q\in[1,\infty)$, we have

$$\|v\|_{q,n} = (\frac{1}{n}\sum_{i=1}^n |v_i|^q)^{1/q}, \quad \|v\|_{\infty,n} = \|v\|_\infty = \max_{i\in[n]}|v_i|.$$

We now define the radii and metric entropies in these norms, with a definition that expands upon Definition 3.1.

Definition E.1 (q-norms, radii, and metric entropies)**.** Given a subset $\mathbb{V}\subset\mathbb{R}^n$ and $q\in[1,\infty]$, define the the *radius* $\mathsf{rad}_q(\mathbb{V}) = \max_{v\in\mathbb{V}}\|v\|_{q,n}$, define the *covering number* $\mathscr{N}(\mathbb{V},\|\cdot\|_{q,n},\varepsilon)$ as the cardinality of minimal-cardinality ε-cover of $\mathbb{V}$ in the norm $\|\cdot\|_{q,n}$, and define the *metric entropy* $\mathscr{M}_q(\mathbb{V},\varepsilon) = \log\mathscr{N}(\mathbb{V},\|\cdot\|_{q,n},\varepsilon)$ as the logarithmic of the covering number. For a function class $\mathcal{H}:\mathcal{W}\to\mathbb{R}$, we define its q-norm *metric entropy* as

$$\mathscr{M}_q(\mathcal{H},\varepsilon) := \sup_{n\in\mathbb{N}}\sup_{w_{1:n}\in\mathcal{W}^n}\mathscr{M}_q(\mathcal{H}[w_{1:n}],\varepsilon).$$

We make the following mild compactness assumption, which holds whenever Theorems 1 and 6 is non-vacuous.

Assumption E.1. For all $\varepsilon>0$, $\mathscr{M}_2(\mathcal{F},\varepsilon)<\infty$.

We also introduce a strictly optional second assumption, but codifies a way in which $\mathcal{F}$ is "simpler" than $\mathcal{G}$, and enables further simplifications when it holds.

Assumption E.2. For all $\varepsilon>0$, $\mathscr{M}_\infty(\mathcal{F},\varepsilon)\le\mathscr{M}_\infty(\mathcal{G},\varepsilon)$.

Lastly, we define the class

$$\beta_{\mathcal{F}} := \{\beta_f : f\in\mathcal{F}\}. \tag{E.1}$$

E.1. An Intermediate Dudley Bound

Recall the Dudley functional from Definition 4.1 We begin by defining an upper bound on the Dudley critical radius $\delta_{n,\mathscr{D}}$ of a function class $\mathcal{H}$

Definition E.2 (Upper Bound on Dudley Critical Radius)**.** We define

$$\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) := \inf_{\delta \le r} \left(2\delta + \frac{4}{\sqrt{n}} \int_{\delta}^{r} \sqrt{\mathscr{M}_2(\mathcal{H}; \varepsilon/4)} \mathrm{d}\varepsilon \right),$$

and

$$\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) := \inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\}.$$

Lastly, we define

$$\bar{\mathscr{D}}_{n,q}(\mathcal{H}) := \inf_{\delta \le B} \left(2\delta + \frac{4}{\sqrt{n}} \int_{\delta}^{B} \sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/2)} \mathrm{d}\varepsilon \right), \quad B := \sup_w |h(w)|.$$

Recall ν_1, ν_2 from Eqs. (3.3) and (3.4). We have the following theorem.

Theorem 5. *Recall the class $\beta_{\mathcal{F}} := \{\beta_f : f \in \mathcal{F}\}$, and suppose Assm. E.1 holds. Then, for any $\delta \in (0,1)$, if $\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, \sigma_B)^2 + \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, \sigma_B)^2 + \frac{\sigma_B^2 \log(1/\delta)}{n} \le c_1 \gamma$. Then it holds that with probability at least $1-\delta$,*

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim (\nu_1+\nu_2)\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, \sigma_B)^2 + \nu_1(\bar{\mathscr{D}}_{n,\infty}(\mathcal{G})^2 + \min\{\bar{\mathscr{D}}_{n,\infty}(\mathcal{F})^2, \bar{\mathscr{D}}_{n,\infty}(\beta_{\mathcal{F}})^2\}) + \nu_2 \cdot \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, \sigma_B)^2 + \frac{(\nu_1+\nu_2)\sigma_B^2 \log(1/\delta)}{n}.$$

In particular, if Assm. E.2 also holds, then

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim (\nu_1+\nu_2)\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, \sigma_B)^2 + \nu_1(\bar{\mathscr{D}}_{n,\infty}(\mathcal{G})^2 + \nu_2 \cdot \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, \sigma_B)^2 + \frac{(\nu_1+\nu_2)\sigma_B^2 \log(1/\delta)}{n}.$$

And in addition, when we upper bound $\nu_1 \le \nu_{x,y}$ and $\nu_2 \le \nu_y$,

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim \nu_{x,y}(\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, \sigma_B)^2 + \bar{\mathscr{D}}_{n,\infty}(\mathcal{G})^2) + \nu_y \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, \sigma_B)^2 + \frac{\nu_{x,y}\sigma_B^2 \log(1/\delta)}{n}.$$

E.1.1. Proof of Theorem 5

We sketch the proof of Theorem 5, deferring supporting proofs to Appendix E.1.2. From Theorem 3, it suffices to establish the following inequalities for $c \ge 1$:

$$\begin{aligned} &\delta_{n,\mathscr{D}}(\mathcal{H}_{\mathtt{sum}}, c) \lesssim \bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, c) + \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, c) \\ &\delta_{n,\mathscr{D}}(\mathcal{F}_{\mathtt{cnt}}, c) \lesssim \bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, c) \\ &\sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{G}_{\mathtt{cnt}}[w_{1:n}]) \lesssim \min\{\bar{\mathscr{D}}_{n,\infty}(\mathcal{F}), \bar{\mathscr{D}}_{n,\infty}(\beta_{\mathcal{F}})\} + \bar{\mathscr{D}}_{n,\infty}(\mathcal{G}). \end{aligned} \tag{E.2}$$

We first upper bound all relevant "non-barred" Dudley integrals in terms of "barred" integrals.

Lemma E.1. *For any class $\mathcal{H}$,*

$$\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) \ge \delta_{n,\mathscr{D}}(\mathcal{H}, c), \quad \textit{and} \quad \bar{\mathscr{D}}_{n,\infty}(\mathcal{H}) \ge \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}]).$$

Next, we give a technical lemma which allows us to relate the Dudley integral/critical radius of class $\mathcal{H}$ in terms of classes which upper bound its metric entropy.

Lemma E.2. *Fix $a \ge 1$.*

(a) Suppose that $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ satisfy the ℓ_2-metric entropy inequality, for all $\varepsilon > 0$,

$$\mathscr{M}_2(\mathcal{H}, \varepsilon) \le \mathscr{M}_2(\mathcal{H}_1, \varepsilon/a) + \mathscr{M}_2(\mathcal{H}_2, \varepsilon/a).$$

Then, it holds that $\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) \le a \max_{i\in[2]} \bar{\delta}_{n,\mathscr{D}}(\mathcal{H}_i, 2c/a)$. In particular, if $a \ge 2$,

$$\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) \le a \max_{i\in[2]} \bar{\delta}_{n,\mathscr{D}}(\mathcal{H}_i, c).$$

(b) Suppose instead $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2$ satisfy the ℓ_∞-metric entropy inequality, for all $\varepsilon > 0$,

$$\mathscr{M}_\infty(\mathcal{H}, \varepsilon) \le \mathscr{M}_\infty(\mathcal{H}_1, \varepsilon/a) + \mathscr{M}_\infty(\mathcal{H}_2, \varepsilon/a).$$

Then, $\bar{\mathscr{D}}_{n,\infty}(\mathcal{H}) \le a\left(\sum_{i=1}^n \bar{\mathscr{D}}_{n,\infty}(\mathcal{H}_i)\right)$

To apply these, we require control over the metric entropy of $\beta_{\mathcal{F}}$.

Lemma E.3. *Let $\beta_{\mathcal{F}} := \{\beta_f : f \in \mathcal{F}\}$. Then, as long as $\mathscr{M}_2(\mathcal{F}, \varepsilon')$ is finite for all ε',*

$$\mathscr{M}_2(\beta_{\mathcal{F}}, \varepsilon) \le \inf_{\varepsilon' < \varepsilon} \mathscr{M}_2(\mathcal{F}, \varepsilon'), \quad \text{and} \quad \mathscr{M}_\infty(\beta_{\mathcal{F}}, \varepsilon) \le \inf_{\varepsilon' < \varepsilon} \mathscr{M}_\infty(\mathcal{F}, \varepsilon')$$

To prove Lemma E.3, we require the following qualitative statement, which can be derived from a Glivenko-Cantelli Theorem (e.g. van der Vaart and Wellner (1996, Theorem 2.8.1), with the substitution $\mathcal{F} \leftarrow (\mathcal{H} - \mathcal{H})^2$).

Proposition E.4 (Uniform Covergence of $\mathcal{L}_2$ measures)**.** *Let P be any measure over $\mathbf{W}$, let $\mathcal{H}$ be any class for which $\mathscr{M}_2(\mathcal{H}, \varepsilon)$ is finite for all ε. Then, for all $t > 0$*

$$\lim_{n\to\infty} \mathbb{P}_{\mathbf{w}_{1:n}\sim P}\left[\sup_{h,h'\in\mathcal{H}} \|h(\mathbf{w}_{1:n}) - h'(\mathbf{w}_{1:n})\|_{2,n} - \mathbb{E}_{\mathbf{w}\sim P}[(h(\mathbf{w}) - h'(\mathbf{w}))^2] \ge t\right] = 0.$$

Proof of Lemma E.3. Fix $w_{1:n} = (x_i, y_i)_{1:n} \in \mathcal{W}^n$ and a slack parameter $t > 0$. Introduce the measure P to be the mixture distribution $P = \frac{1}{n}\sum_{i=1}^n \mathbb{P}_{\mathsf{train}}[\mathbf{x} = \cdot \mid \mathbf{y} = y_i]$. Recall $\mathcal{H}_2 = \{\beta_f : f \in \mathcal{F}\}$,

$$\begin{aligned}
\|\beta_f[w_{1:n}] - \beta_{f'}[w_{1:n}]\|_{2,n}^2 &= \sum_{i=1}^n (\mathbb{E}_{\mathsf{train}}[f(\mathbf{x}) - f'(\mathbf{x}) \mid \mathbf{y} = y_i])^2 \\
&\le \sum_{i=1}^n \mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2 \mid \mathbf{y} = y_i] \\
&= \mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2]
\end{aligned}$$

Proposition E.4 implies that there must exists some number $m \ge n$ and $\tilde{x}_{1:m} \in \mathcal{X}^m$ such that, for all $f, f' \in \mathcal{F}$,

$$\mathbb{E}_{\mathbf{x}\sim P}\mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2] \le t + \|f[\tilde{x}_{1:m}] - f'[\tilde{x}_{1:m}]\|_{2,m}^2$$

Hence,

$$\mathscr{M}_2(\mathcal{H}_2[w_{1:n}], \varepsilon + t) \le \mathscr{M}_2(\mathcal{F}[\tilde{x}_{1:m}], \varepsilon) \le \mathscr{M}_2, \varepsilon).$$

As $t, w_{1:n}$ are arbitrary, the first bound follows. To prove the second, we apply a similar argument, bounding

$$\|\beta_f[w_{1:n}] - \beta_{f'}[w_{1:n}]\|_\infty^2 \le \max_{i\in[n]} \mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2 \mid y = y_i],$$

For each y_i, there exists some m_i and a sequence $\tilde{x}_{1:m_i}^{(i)}$ such that for all $f, f' \in \mathcal{F}$,

$$\mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2 \mid \mathbf{y} = y_i] \le t + \|f[\tilde{x}_{1:m}] - f'[\tilde{x}_{1:m}]\|_{2,m}^2 \le t \max_{j\in[m_i]} |\tilde{x}_j^{(i)}] - f'[\tilde{x}_j^{(i)}]|_\infty^2$$

Hence, introduce the finite set $\tilde{\mathcal{X}} := \bigcup_{i=1}^n \bigcup_{j=1}^{m_i} \{x_j^{(i)}\}$, we have that for all $f, f' \in \mathcal{F}$,

$$\max_i \mathbb{E}_{\mathsf{train}}[(f(\mathbf{x}) - f'(\mathbf{x}))^2 \mid \mathbf{y} = y_i] \le t + \max_{\tilde{x}\in\tilde{\mathcal{X}}} |f(\tilde{x}) - f'(\tilde{x})|^2.$$

The bound follows. □

Eq. (E.2) is now an immediate consequence of Lemma E.1 and the following lemma.

Lemma E.5. *The following bounds hold:*

(a) $\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}_{\texttt{sum}}, c) \le 2\max\{\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, c), \bar{\delta}_{n,\mathscr{D}}(\mathcal{G}, c)\}$

(b) $\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}_{\texttt{cnt}}, c) \le 4\bar{\delta}_{n,\mathscr{D}}(\mathcal{F}, c)$.

(c) $\bar{\mathscr{D}}_{n,\infty}(\mathcal{G}_{\texttt{cnt}}) \le \min\{\bar{\mathscr{D}}_{n,\infty}(\mathcal{F}), \bar{\mathscr{D}}_{n,\infty}(\beta_{\mathcal{F}})\} + \bar{\mathscr{D}}_{n,\infty}(\mathcal{G})$.

Proof of Lemma E.5. For all points, we apply Lemma E.2. For (a), we can verify by the triangle inequality that for any $\mathcal{H}_1, \mathcal{H}_2$,

$$\mathscr{M}_2(\mathcal{H}_1 + \mathcal{H}_2, \varepsilon) \le \mathscr{M}_2(\mathcal{H}_1, \varepsilon/2) + \mathscr{M}_2(\mathcal{H}_2, \varepsilon/2), \tag{E.3}$$

which yields part (a) when specializing to $\mathcal{H}_1 = \mathcal{F}$, $\mathcal{H}_2 = \mathcal{G}$, and applying Lemma E.2. For part (b), we observe that $\mathcal{F}_{\texttt{cnt}} \subset (\mathcal{F} - f_\star) - \beta_{\mathcal{F}}$. Thus, Fact E.1, followed by Eq. (E.3) and finally Lemma E.3 imply

$$\begin{aligned}
\mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon) &\le \mathscr{M}_2(\mathcal{F}_{\texttt{cnt}} - \beta_{\mathcal{F}}, \varepsilon/2) && \text{(Fact E.1)} \\
&\le \mathscr{M}_2(\mathcal{F} - f_\star, \varepsilon/4) + \mathscr{M}_2(\beta_{\mathcal{F}}, \varepsilon/4) && \text{(Eq. (E.3))} \\
&= \mathscr{M}_2(\mathcal{F}, \varepsilon/4) + \mathscr{M}_2(\beta_{\mathcal{F}}, \varepsilon/4) \\
&\le \mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon/4) + \inf_{b>1} \mathscr{M}_2(\mathcal{F}, \varepsilon/4b) && \text{(Lemma E.3)} \\
&\le \inf_{b>1} \mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon/4b) + \mathscr{M}_2(\mathcal{F}, \varepsilon/4b).
\end{aligned}$$

The result now follows from Lemma E.2 with $a \leftarrow 4b$, and taking $b \to 1$.

The proof of part (c) is similar:

$$\begin{aligned}
\mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon) &= \mathscr{M}_2(\mathcal{G} - g_\star + \beta_{\mathcal{F}}, \varepsilon) && \text{(Fact E.1)} \\
&\le \mathscr{M}_2(\mathcal{G} - g_\star, \varepsilon/2) + \mathscr{M}_2(\beta_{\mathcal{F}}, \varepsilon/2) && \text{(Eq. (E.3))} \\
&= \mathscr{M}_2(\mathcal{G}, \varepsilon/2) + \mathscr{M}_2(\beta_{\mathcal{F}}, \varepsilon/2) \\
&\le \mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon/2) + \inf_{b>1} \mathscr{M}_2(\mathcal{F}, \varepsilon/2b) && \text{(Lemma E.3)} \\
&\le \inf_{b>1} \mathscr{M}_2(\mathcal{F}_{\texttt{cnt}}, \varepsilon/4b) + \mathscr{M}_2(\mathcal{F}, \varepsilon/2b),
\end{aligned}$$

and follows from similar steps as part (c). □

E.1.2. Proofs of supporting Lemmas for Theorem 5

Fact E.1. [Exercise 4.2.10 in (Vershynin, 2018)] For any sets $\mathbb{V}' \subset \mathbb{V}$, ε-covering number of $\mathbb{V}'$ in any norm is at most the $\varepsilon/2$ covering number of $\mathbb{V}$.

Proof of Lemma E.1. The proof of $\bar{\mathscr{D}}_{n,\infty}(\mathcal{H}) \ge \sup_{w_{1:n}} \mathscr{D}_{n,\infty}(\mathcal{H}[w_{1:n}])$. is straightforward. To check $\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) \ge \delta_{n,\mathscr{D}}(\mathcal{H}, c)$, we invoke Fact E.1:

$$\sup_{w_{1:n}} \mathscr{M}_q(\mathcal{H}[r, w_{1:n}]; \varepsilon/2) \le \sup_{w_{1:n}} \mathscr{M}_q(\mathcal{H}[w_{1:n}]; \varepsilon/4) \le \mathscr{M}_q(\mathcal{H}, \varepsilon/4). \tag{E.4}$$

Thus,

$$\begin{aligned}
\sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) &:= \sup_{w_{1:n}} \inf_{\delta \le R} \left(2\delta + \frac{4}{\sqrt{n}} \int_\delta^R \sqrt{\mathscr{M}_q(\mathcal{H}[r, w_{1:n}]; \varepsilon/2)} \mathrm{d}\varepsilon\right), \quad \text{where } R = \mathsf{rad}_2(\mathcal{H}[r, w_{1:n}]) \\
&\le \sup_{w_{1:n}} \inf_{\delta \le r} \left(2\delta + \frac{4}{\sqrt{n}} \int_\delta^r \sqrt{\mathscr{M}_q(\mathcal{H}[r, w_{1:n}]; \varepsilon/2)} \mathrm{d}\varepsilon\right) \quad (\mathsf{rad}_2(\mathcal{H}[r, w_{1:n}] \le r \text{ by localization}) \\
&\le \inf_{\delta \le r} \left(2\delta + \frac{4}{\sqrt{n}} \int_\delta^r \sup_{w_{1:n}} \sqrt{\mathscr{M}_q(\mathcal{H}[r, w_{1:n}]; \varepsilon/2)} \mathrm{d}\varepsilon\right) \\
&\le \inf_{\delta \le r} \left(2\delta + \frac{4}{\sqrt{n}} \int_\delta^r \sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/4)} \mathrm{d}\varepsilon\right) && \text{(Eq. (E.4))} \\
&:= \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r).
\end{aligned}$$

Hence,

$$\begin{aligned}\delta_{n,\mathscr{D}}(\mathcal{H}, c) &:= \inf\left\{r : \sup_{w_{1:n}} \mathscr{D}_{n,2}(\mathcal{H}[r, w_{1:n}]) \le \frac{r^2}{2c}\right\} \\ &\le \inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\} = \bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c).\end{aligned}$$ □

Lemma E.6 (Concavity of $\bar{\mathscr{D}}$)**.** *For all $a \ge 1$, $\bar{\mathscr{D}}_{n,2}(\mathcal{H}, ar) \le a\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r)$. Hence, if $\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}$, then $\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r') \le \frac{(r')^2}{2c}$ for $r' \ge r$. Moreover, if $b \ge 1$, $\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, br) \le b\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, r)$.*

Proof of Lemma E.6.

$$\begin{aligned}\bar{\mathscr{D}}_{n,2}(\mathcal{H}, ar) &= \inf_{\delta \le ar}\left(2\delta + \frac{4}{\sqrt{n}}\int_\delta^{ar}\sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &= \inf_{\delta \le ar}\left(2\delta + \frac{4a}{\sqrt{n}}\int_{\delta/a}^{r}\sqrt{\mathscr{M}_q(\mathcal{H}; a\varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &\le \inf_{\delta \le ar}\left(2\delta + \frac{4a}{\sqrt{n}}\int_{\delta/a}^{r}\sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &\le \inf_{\delta \le r}\left(2a\delta + \frac{4a}{\sqrt{n}}\int_{\delta}^{r}\sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/4)}\mathrm{d}\varepsilon\right) = a\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r).\end{aligned}$$

The rest of the result follows similarly to Lemma C.3. □

Proof of Lemma E.2. We prove part (a), the calculate for part (b) is near-identical.

$$\begin{aligned}\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) &\le \inf_{\delta \le r}\left(2\delta + \frac{4}{\sqrt{n}}\int_\delta^{r}\sqrt{\mathscr{M}_q(\mathcal{H}; \varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &\le \inf_{\delta \le r}\left(2\delta + \frac{4}{\sqrt{n}}\sum_{i=1}^{2}\int_\delta^{r}\sqrt{\mathscr{M}_q(\mathcal{H}_i; \varepsilon/4a)}\mathrm{d}\varepsilon\right) && \text{(by assumption)} \\ &\le \inf_{\delta \le r}\left(2\delta + \frac{4a}{\sqrt{n}}\sum_{i=1}^{2}\int_{\delta/a}^{r/a}\sqrt{\mathscr{M}_q(\mathcal{H}_i; \varepsilon/4)}\mathrm{d}\varepsilon\right) && (\varepsilon \leftarrow \varepsilon/a) \\ &= a\inf_{\delta \le r/a}\left(2\delta + \frac{4}{\sqrt{n}}\sum_{i=1}^{2}\int_{\delta}^{r/a}\sqrt{\mathscr{M}_q(\mathcal{H}_i; \varepsilon/4)}\mathrm{d}\varepsilon\right) && (\delta \leftarrow a\delta) \\ &= a\inf_{\delta_1, \delta_2 \le r/a}\left(2\max\{\delta_1, \delta_2\} + \frac{4}{\sqrt{n}}\sum_{i=1}^{2}\int_{\max\{\delta_1,\delta_2\}}^{r/a}\sqrt{\mathscr{M}_q(\mathcal{H}_i; \varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &\le a\inf_{\delta_1, \delta_2 \le r}\sum_{i=1}^{2}\left(2\delta_i + \frac{4}{\sqrt{n}}\int_{\delta_i}^{r/a}\sqrt{\mathscr{M}_q(\mathcal{H}_i; \varepsilon/4)}\mathrm{d}\varepsilon\right) \\ &= a\sum_{i=1}^{2}\bar{\mathscr{D}}_{n,2}(\mathcal{H}_i, \frac{r}{a}).\end{aligned}$$

Consequently,

$$\begin{aligned}\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) &:= \inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\} \\ &\le \inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}_1, \frac{r}{a}) + \bar{\mathscr{D}}_{n,2}(\mathcal{H}_2, \frac{r}{a}) \le \frac{r^2}{2ac}\} \\ &= a\inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}_1, r) + \bar{\mathscr{D}}_{n,2}(\mathcal{H}_2, r) \le \frac{ar^2}{2c}\} \\ &= a\inf\{r : \max_{i\in[2]} \bar{\mathscr{D}}_{n,2}(\mathcal{H}_1, r) \le \frac{ar^2}{4c}\}.\end{aligned}$$

By Lemma E.6, the above is at most $a\max_{i\in[2]}\inf\{r : \bar{\mathscr{D}}_{n,2}(\mathcal{H}_1, r) \le \frac{r^2}{2c}\} \le a\max_{i\in[2]}\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}_i, 2c/a)$. □

E.2. Instantiating the Rates

Next, we formally define families of function classes we all *entropy families*, which are characterized by upper bounds of their metric entropies. These entropy families formally capture the entropy rates depicted in Theorem 1.

Definition E.3 (Entropy Families)**.** Let $\mathcal{H} : \mathcal{W} \to \mathbb{R}$, and let $q \in [1, \infty]$, and let $\boldsymbol{\tau} = (\tau_0, \tau_1, \tau_2) \in \mathbb{R}^3$ denote a vector of parameters. We say that $\mathcal{H} \in \mathsf{EntFam}_q(p, \boldsymbol{\tau}; R)$ if $\mathsf{rad}_q(\mathcal{H}) \le R$ and either

- $p = 0$, and for all $\varepsilon > 0$, $\mathscr{M}_q(\mathcal{H}) \le \tau_0 + \tau_1\log(\tau_2/\varepsilon)$ or
- $p > 0$, and for all $\varepsilon > 0$, $\mathscr{M}_q(\mathcal{H}, \varepsilon) \le \tau_0 + \tau_1\varepsilon^{-p}$.

Notice that the sets $\mathsf{EntFam}_q(p, \boldsymbol{\tau})$ are non-increasing in q, and non-decreasing in the coordinates of $\boldsymbol{\tau}$, and (up to constants) non-increasing in p.

We now define complexities measures that upper bound the localized and unlocalized Dudley integrals for function classes in a given entropy family.

Definition E.4 (Key Complexities)**.** Let $\boldsymbol{\tau} = (\tau_0, \tau_1, \tau_2) \in \mathbb{R}^3_{\ge 0}$, $p \in [0, \infty)$, and $R, c > 0$. We define the *global complexity term*

$$\Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}, R) := \frac{R\tau_0}{\sqrt{n}} + \begin{cases} R\sqrt{\frac{\tau_1\log(e+\tau_2/R)}{n}} & p = 0 \\ \frac{R^{1-p/2}}{1-p/2}\sqrt{\frac{\tau_1}{n}} & p \in (0,2) \\ \sqrt{\frac{\tau_1}{n}}\log(e + R\sqrt{n/\tau_1}) & p = 2 \\ (\frac{\tau_1}{n})^{\frac{1}{p}}(p/2-1)^{-\frac{2}{p}} & p > 2 \end{cases} = \widetilde{\mathcal{O}}(1)\cdot\begin{cases} n^{-\frac{1}{2}} & p \le 2 \\ n^{-\frac{1}{p}} & p > 2 \end{cases}$$

and the *local complexity term*

$$\Xi_{n,\mathtt{loc}}(p, \boldsymbol{\tau}, c) := \frac{c^2(1+\tau_0)^2}{n} + \begin{cases} \frac{c^2\tau_1\log(e+\sqrt{n}\tau_2/c)}{n} & p = 0 \\ \left(\frac{c^2}{(1-p/2)^2}\cdot\frac{\tau_1}{n}\right)^{\frac{2}{2+p}} & p \in (0,2) \\ c\sqrt{\frac{\tau_1\log(e+c\sqrt{n}/\tau_1)}{n}} & p = 2 \\ (p/2-1)^{-\frac{2}{p}}(\frac{\tau_1}{n})^{\frac{1}{p}} & p > 2. \end{cases} = \widetilde{\mathcal{O}}(1)\cdot\begin{cases} n^{-\frac{2}{2+p}} & p \le 2 \\ n^{-\frac{1}{p}} & p > 2 \end{cases}$$

Lastly, we define the *rate functionals*.

Definition E.5 (Rate Functionals)**.** We define the following rate functionals:

$$\begin{aligned}\mathrm{rate}_{n,q}(\mathcal{H}, c) &:= \inf_{p,\boldsymbol{\tau},R}\{\Xi_{n,\mathtt{loc}}(p, \boldsymbol{\tau}, c) : \mathcal{H} \in \mathsf{EntFam}_q(p, \boldsymbol{\tau}; R)\} \\ \mathrm{rate}_{n,\star}(\mathcal{H}) &:= \inf_{p,\boldsymbol{\tau},R}\{\Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}, R) : \mathcal{H} \in \mathsf{EntFam}_\infty(p, \boldsymbol{\tau}; R)\}.\end{aligned}$$

That is, $\mathrm{rate}_{n,q}(\mathcal{H}, c)$ is the smallest possibly local complexity term subject to $\mathcal{H}$ being in the appropriate entropy family, and $\mathrm{rate}_{n,\star}(\mathcal{H})$ is the smallest possible global complexity term, always taken with metric entropy in the ∞-norm. We are now ready to state our main theorem with explicit rates:

Theorem 6. *Suppose Assms. 2.1 to 2.3 hold. Let $\sigma_B := \max\{B, \sigma\}$, let ν_1, ν_2 be as in Eqs. (3.3) and (3.4), and let c_2 be a sufficiently small universal constant. Then if*

$$\mathrm{rate}_{n,2}(\mathcal{G}, \sigma_B) + \mathrm{rate}_{n,2}(\mathcal{F}, \sigma_B) + \frac{\sigma_B^2 \log(1/\delta)}{n} \le c_2\gamma,$$

it holds that probability at least $1-\delta$, (recalling the class $\beta_{\mathcal{F}} := \{\beta_f : f \in \mathcal{F}\}$),

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim (\nu_1+\nu_2)\mathrm{rate}_{n,2}(\mathcal{F}, \sigma_B) + \nu_1\left(\mathrm{rate}_{n,\star}(\mathcal{G})^2 + \min\{\mathrm{rate}_{n,\star}(\mathcal{F})^2, \mathrm{rate}_{n,\star}(\beta_{\mathcal{F}})^2\}\right) + \nu_2\,\mathrm{rate}_{n,2}(\mathcal{G}, \sigma_B) + (\nu_1+\nu_2)\cdot\frac{\sigma_B^2\log(1/\delta)}{n}.$$

If in addition Assm. E.2 holds, obtain

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim (\nu_1+\nu_2)\mathrm{rate}_{n,2}(\mathcal{F}, \sigma_B) + \nu_1\mathrm{rate}_{n,\star}(\mathcal{G})^2 + \nu_2\,\mathrm{rate}_{n,2}(\mathcal{G}, \sigma_B) + (\nu_1+\nu_2)\cdot\frac{\sigma_B^2\log(1/\delta)}{n},$$

and in addition, if we upper bound $\nu_1 \le \nu_{x,y}$ and $\nu_2 \le \nu_y$, we obtain (as $\nu_y \le \nu_{x,y}$)

$$\mathcal{R}_{\mathsf{test}}(\hat{f}_n, \hat{g}_n) \lesssim \nu_{x,y}\left(\mathrm{rate}_{n,2}(\mathcal{F}, \sigma_B) + \mathrm{rate}_{n,\star}(\mathcal{G})^2\right) + \nu_y\,\mathrm{rate}_{n,2}(\mathcal{G}, \sigma_B) + \nu_{x,y}\frac{\sigma_B^2\log(1/\delta)}{n}.$$

Proof. The proof is a direct consequence of Theorem 3 and the following two lemmas to bound the Dudley integrals and critical radii, whose computations are essentially standard but which we prove in the Appendix E.3. □

Lemma E.7. *Suppose that $\mathcal{H} \in \mathsf{EntFam}_q(p, \boldsymbol{\tau}, R)$. Then, $\bar{\mathscr{D}}_{n,q}(\mathcal{H}) \lesssim \Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}; R)$. Hence,*

$$\bar{\mathscr{D}}_{n,\infty}(\mathcal{H}) \le \mathrm{rate}_{n,\star}(\mathcal{H}).$$

Lemma E.8. *Suppose that $\mathcal{H} \in \mathsf{EntFam}_2(p, \boldsymbol{\tau}, R)$, and let $c > 0$ be arbitrary. Then, $\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c)^2 \lesssim \Xi_{n,\mathtt{loc}}(p, \boldsymbol{\tau}, c)$. Hence,*

$$\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c)^2 \le \mathrm{rate}_{n,2}(\mathcal{H}, c).$$

E.3. Proof of Dudley Bounds (Lemmas E.7 and E.8)

Proof of Lemma E.7. From the definition of the Dudley functional, Definition 4.1, it is clear that the additive τ_0-term in the metric entric bound contributes at most an additive $\frac{R\tau_0}{\sqrt{n}}$ term to the integral. Consequently, let handle what is left over, assuming throughout that $\tau_0 = 0$. For a class $\mathcal{H}$, let $\phi(\varepsilon)$ be the upper bound on $\mathscr{M}_q(\mathcal{H}, \varepsilon)$ prescribed by Definition E.3 (again, setting $\tau_0 = 0$). Then, from Definition 4.1, so that

$$\bar{\mathscr{D}}_{n,q}(\mathcal{H}) \le \inf_{\delta \le R}\left(2\delta + \frac{4}{\sqrt{n}}\int_\delta^R \sqrt{\phi(\varepsilon/2)}\mathrm{d}\varepsilon\right) \lesssim \inf_{\delta \le R}\left(\delta + \frac{1}{\sqrt{n}}\int_\delta^R \sqrt{\phi(\varepsilon)}\mathrm{d}\varepsilon\right),$$

where last inequality uses similar changes of variables as in Lemma E.2. When $p < 2$, we take $\delta = 0$ and attain

$$\bar{\mathscr{D}}_{n,q}(\mathcal{H}) \lesssim R\frac{1}{\sqrt{n}}\int_0^R \sqrt{\phi(\varepsilon)}\mathrm{d}\varepsilon \lesssim R\frac{\tau_0}{\sqrt{n}} + R\sqrt{\frac{\tau_1}{n}}\cdot\begin{cases}\sqrt{\log(\tau_2 R)} & p = 0\\ \frac{1}{p/2-1}R^{1-p/2} & p \in (0,2)\end{cases}.$$

Next, for the case $p = 2$, we pick $\delta = R\sqrt{\tau_1/n}$ to get

$$\inf_{\delta \le R}\delta + \frac{1}{\sqrt{n}}\int_\delta^R \sqrt{\phi(\varepsilon)} \le R\frac{\tau_0}{\sqrt{n}} + \inf_{\delta \le R}\delta + \sqrt{\tau_1/n}\log(R/\delta) \lesssim R\frac{\tau_0}{\sqrt{n}} + \sqrt{\tau_1/n}\log(e + R\sqrt{n/\tau_1}).$$

Finally, consider $p > 2$. Then,

$$\inf_{\delta \le R} \delta + \frac{1}{\sqrt{n}} \int_{\delta}^{R} \sqrt{\phi(\varepsilon)} \le R\frac{\tau_0}{\sqrt{n}} + \inf_{\delta \le R} \delta + \frac{1}{p/2-1}\sqrt{\tau_1/n}\delta^{1-p/2}$$

Taking $\delta = \frac{1}{p/2-1}\sqrt{\tau_1}\delta^{1-p/2}$, we choose $\delta = \tau_1^{1/p}(p/2-1)^{-2/p}$ yielding

$$\inf_{\delta \le R} \delta + \frac{1}{\sqrt{n}} \int_{\delta}^{R} \sqrt{\phi(\varepsilon)} \lesssim R\frac{\tau_0}{\sqrt{n}} + (\tau_1/n)^{1/p}(p/2-1)^{-2/p}.$$

This concludes the proof. □

Proof of Lemma E.8. Modifying the computation in Lemma E.7 implies

$$\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \lesssim \Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}; r).$$

. For $p = 0$, we use that that for $r \ge c/\sqrt{n}$,

$$\Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}; r) = r\left(\frac{\tau_0 + \sqrt{\tau_1 \log(e + \tau_2/r)}}{\sqrt{n}}\right) \le r\left(\frac{(e+\tau_0) + \sqrt{\tau_1 \log(e + \sqrt{n}\tau_2/c)}}{\sqrt{n}}\right),$$

so that

$$\bar{\delta}_{n,\mathscr{D}}(\mathcal{H}, c) = \inf\left\{r^2 : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\right\} \lesssim c^2\left(\frac{\tau_0^2 + \tau_1 \log(e + \sqrt{n}\tau_2/c)}{n}\right) := \Xi_{n,\mathtt{loc}}(0, \boldsymbol{\tau}; c)$$

For $p \in (0, 2)$,

$$\Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}; r) = \frac{r\tau_0}{\sqrt{n}} + \frac{r^{1-p/2}}{1-p/2}\sqrt{\frac{\tau_1}{n}}$$

Going forward, let c_0 a universal constant for which $\bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le c_0 \Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}; r)$. Thus, one can check that

$$\inf\left\{r^2 : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\right\} \lesssim \max\{r_1^2, r_2^2\},$$

where r_1 and r_2 balance the following equations

$$\frac{r_1(1+\tau_0)}{\sqrt{n}} = r_1^2/4c_0c, \quad \frac{r_2^{1-p/2}}{1-p/2}\sqrt{\frac{\tau_1}{n}} = r_2^2/4c_0c.$$

Solving yields

$$r_1^2 = \frac{16c_0^2c^2(1+\tau_0)^2}{n}, \quad r_2^2 = \left(\frac{4c_0c}{1-p/2}\sqrt{\frac{\tau_1}{n}}\right)^{\frac{2}{1+p/2}} = \left(\frac{1}{(1-p/2)^2} \cdot \frac{16c_0^2c^2\tau_1}{n}\right)^{\frac{2}{2+p}}$$

which gives

$$\max\{r_1^2, r_2^2\} \lesssim \frac{c^2(1+\tau_0)^2}{n} + \left(\frac{c^2}{(1-p/2)^2} \cdot \frac{\tau_1}{n}\right)^{\frac{2}{2+p}} := \Xi_{n,\mathtt{loc}}(p, \boldsymbol{\tau}; c), \quad p \in (0, 2).$$

For $p = 2$, we note that

$$\Xi_{n,\mathtt{glob}}(p, \boldsymbol{\tau}, \tau_1; r) \lesssim \left(\frac{r\tau_0 + \sqrt{\tau_1 \log(e + R\sqrt{n}/\tau_1)}}{\sqrt{n}}\right).$$

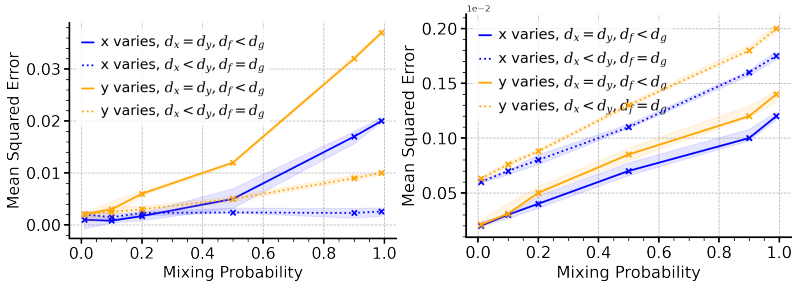

Figure 2. **Mean Squared Error of predictors** of form $\hat{z} = f_\theta(x) + g_\theta(y)$ (Left) and $\hat{z} = h_\theta(x, y)$ (Right) on shifted distributions, where we hold the mixing probability of one of $\{\mathbf{x}, \mathbf{y}\}$ fixed, and vary the other's probability. MSE for both predictors declines less with shifts in p_x than those in p_y.

So similarly,

$$\inf\left\{r^2 : \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\right\} \lesssim \max\{r_1^2, r_3^2\},$$

where r_1 is as above, and where r_3 is the smallest term satisfying

$$\frac{\sqrt{\tau_1 \log(e + r_3\sqrt{n}/\tau_1)}}{\sqrt{n}} \le r_3^2/4c_0c$$

As $\frac{\sqrt{\tau_1 \log(e+r_3\sqrt{n}/\tau_1)}}{\sqrt{n}} \le r_3$ by Jensen's inequality, we can take $r_3 \le 4c_0c$. Thus, it suffices that r_3 satisfy

$$r_3 \le 4c_0c\frac{\sqrt{\tau_1 \log(e + 4c_0c\sqrt{n}/\tau_1)}}{\sqrt{n}},$$

so that suppressing the universal constant c_0,

$$r_1^2 + r_3^2 \lesssim \frac{c^2(1+\tau_0^2)}{n} + c\frac{\sqrt{\tau_1 \log(e + c\sqrt{n}/\tau_1)}}{\sqrt{n}} = \Xi_{n,\texttt{loc}}(2, \boldsymbol{\tau}; c).$$

For $p > 2$, repeating the same arguments as above, we have

$$\inf\left\{r^2 : \sup_{w_{1:n}} \bar{\mathscr{D}}_{n,2}(\mathcal{H}, r) \le \frac{r^2}{2c}\right\} \lesssim \max\{r_1^2, r_4^2\},$$

where r_1 is as above and r_4 satisfies

$$r_4^2/4c_0c = (\frac{\tau_1}{n})^{1/p}(p/2-1)^{-\frac{2}{p}}.$$

so that $r_4^2 \lesssim c(\frac{\tau_1}{n})^{1/p}(p/2-1)^{-\frac{2}{p}}$. Following similar steps concludes the proof. □

F. Experiment Details

This section describes various experiment details regarding model architectures and training hyperparameters.

F.1. Regression

Analyzing *simple* feature To make $\mathbf{x}$ simpler, we either have $d_x < d_y$ or $d_f < d_g$. We compare the generalization error (i.e., the difference between test and train mean squared error) for the auxiliary task of predicting $f_\star(\mathbf{x})$ and $g_\star(\mathbf{y})$. We train $f_\phi(\mathbf{x})$ to predict $f_\star(\mathbf{x})$ and $g_\phi(\mathbf{y})$ to predict $g_\star(\mathbf{y})$. f_ϕ incurs a generalization error of 2.5×10^{-5}. When $d_x < d_y$, g_ϕ incurs a generalization error of 4.9×10^{-5}. When $d_f < d_g$, g_ϕ incurs a generalization error of 7.8×10^{-5}. In both cases, we observe that the auxiliary task of predicting $f_\star(\mathbf{x})$ has less generalization error.

Testing resiliency of predictors We independently train two predictors $\hat{z} = f_\theta(x) + g_\theta(y)$ and $\hat{z} = h_\theta(x, y)$ (using concatenated features $(\mathbf{x}, \mathbf{y})$) to minimize mean-square error (MSE) under a training distribution with $p_x = p_y = .01$. We then measure the MSE for both predictors on shifted distributions, where we hold the mixing probability of one of $\{\mathbf{x}, \mathbf{y}\}$ fixed, and vary the other's probability in the range $\{0.1, 0.2, 0.5, 0.9, 0.99\}$. Figure 2 shows that MSE for both predictors declines less with shift in p_x than with those in p_y, thereby corroborating our theoretical expectations.

Implementation details We use $d_x = 3$ and d_y is either 3 (when $d_x = d_y$) or 6 (when $d_x < d_y$). Both $f(x)$ and $g(y)$ are 2-layered Multi-layered perceptions (MLP) with ReLU activation. While f has a hidden dimension $d_f = 32$, g has a hidden dimension of either $d_g = 32$ (when $d_f = d_g$) or $d_g = 4096$ (when $d_f < d_g$). We parameterize $f_\theta(x)$ and $f_\phi(x)$ with 2-layered MLPs having ReLU activation and same hidden dimension as f. We parameterize $g_\theta(y)$ and $g_\phi(y)$ with 2-layered MLPs having ReLU activation and same hidden dimension as g. We parameterize $h_\theta(x, y)$ with a 2-layered MLP having ReLU activation and same hidden dimension as g. We collect $100k$ data points with $p_x = p_y = 0.01$ for training h_θ. We train h_θ with Adam optimizer (Kingma and Ba, 2014) for 100 epochs using a learning rate of 0.001 and a batch size of 50. During test, we increase p_x and p_y to $\{0.1, 0.2, 0.5, 0.9, 0.99\}$ and evaluate the model using 1000 data points.

F.2. Binary Classification with Waterbird dataset

Setup Our task is to classify images of birds as waterbirds or landbirds, against a background of either land or water. These images have two high-level features: `birdtype` and `background`. We first empirically determine which feature is *simple*. We then empirically test if the classifier is more resilient to distribution shift in the simple feature.

Determining *simple* feature To determine which feature is simple, we learn two classifiers, f_{bird} predicting `birdtype` and f_{back} predicting `background`. We train them on standard training set of waterbird dataset and test them on a sampled test set that has the same distribution as the training set. While both f_{bird} and f_{back} have a training accuracy of 1.0, f_{bird} has a test accuracy of 0.88 and f_{back} has a test accuracy of 0.96. Since f_{back} has lower generalization error, we consider `background` as simple.

Testing classifier resiliency We test resiliency of f_{bird} as we shift distribution of one feature while keeping the distribution of the other feature fixed. Specifically, we vary the proportion (in percentages) of images with waterbird (resp. land background) in test set while keeping the proportion of images with land background (resp. waterbird) fixed. We show the results in Figure 1. We observe that test accuracy of f_{bird} varies less when we shift distribution of `background` while keeping the distribution of `birdtype` fixed.

Implementation details We parameterize f_{bird} and f_{back} with ResNet50 model (He et al., 2016) and train them with stochastic gradient descent for 300 epochs using a learning rate of 0.001, batch size of 128, momentum of 0.9 and l2 regularization of 0.0001. We took these hyperparameters and architectural choices from Sagawa et al. (2019) which introduced the Waterbird dataset. We used the github repo https://github.com/kohpangwei/group_DRO for running our experiments.

F.3. Multi-class Classification with FMoW

Setup Our task is to predict `landtype` from images with top down satellite view. These images also contain information about their geographical region (Africa, the Americas, Oceania, Asia, or Europe). Hence, they have two high-level features: `landtype` and `region`. We first empirically determine which feature is *simple*. We then empirically test if the classifier is more resilient to distribution shift in the simple feature.

Determining *simple* feature To determine which feature is simple, we learn two classifiers, f_{land} predicting `landtype` and f_{geo} predicting `region`. We train and test them on standard training and test set of FMoW dataset. While f_{land} and f_{geo} have a training accuracy (i.e. number of correct predictions/ number of datapoints) of 0.71 and 0.74, they have a test accuracy of 0.59 and 0.69 respectively. Since f_{geo} has lower generalization error, we consider `region` as simple.

Testing classifier resiliency We test resiliency of f_{land} as we shift distribution of one feature while keeping the distribution of the other feature fixed. To shift the distribution of `landtype`, we vary the proportion of images with labelled as zoo (a `landtype`). Similarly, to shift the distribution of `region`, we vary the proportion of images from `Africa` region. We

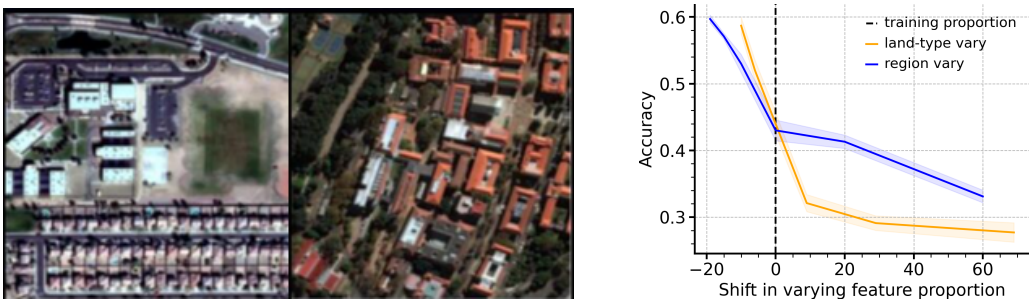

Figure 3. (Left) **Visualization of FMoW dataset**. (Right) **Test accuracy of `landtype` classifier** as we shift distribution of `region` (resp. `landtype`) while keeping the distribution of `landtype` (resp. `region`) fixed. The results again show that the classifier is more robust to distribution shift in simple feature `region`.

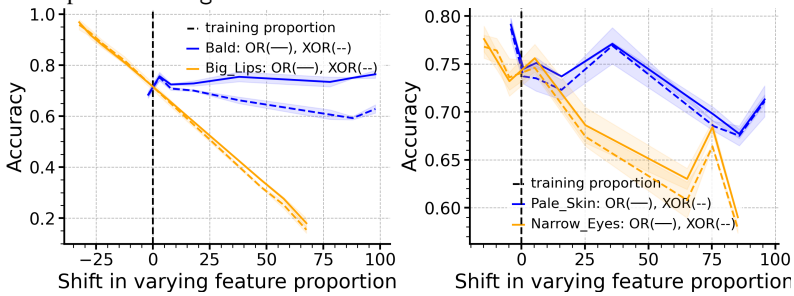

Figure 4. **Test accuracy of logical operators over a pair of (*simple*, *complex*) attribute**. We vary the proportion (in percentages) of images with simple attribute (resp. complex attribute) in test set while keeping the proportion of images with complex attribute (resp. simple attribute) fixed. We use pairs {`bald`, `big_lips`} (Left) and {`pale_skin`, `narrow_eyes`} (Right).

show the results in Figure 3. We observe that test accuracy of f_{land} varies less (i.e. f_{land} is more resilient) when we shift distribution of `region` while keeping the distribution of `landtype` fixed.

Implementation details We parameterize f_{land} and f_{geo} with DenseNet121 model (Huang et al., 2017) and train them with Adam optimizer (Kingma and Ba, 2014) for 60 epochs using a learning rate of 0.0001 and batch size of 32. We took these hyperparameters and architectural choices from Koh et al. (2021). We used the github repo https://github.com/p-lambda/wilds for running our experiments.

F.4. Learning logical operators with CelebA

Setup We re-purpose the CelebA dataset to learn logical operators OR and XOR for two attributes. We first empirically determine which attributes are *simpler* than others. We then learn logical operators combining a *simple* attribute and a *complex* attribute. Finally, we empirically test the resilience of logical operators against distribution shifts in simple and complex attributes.

Determining *simple* attributes We first train and test a multi-head binary classifier that detects presence of 40 different attributes, with one head per attribute, on images from the CelebA "standard training set" (CelebA-STS). We select attributes {`bald`, `pale_skin`} as "simple" due to their low generalization error, and {`big_lips`, `narrow_eyes`} as "complex" due to their larger generalization error. Table 1 shows a complete list of training and test accuracy of the multi-head binary classifier for each attribute.

Testing resiliency of logical operators We first learn logical operators f_{OR} and f_{XOR} over a pair of simple and complex attribute. We use two such pairs in our experiments: {`bald`, `big_lips`} and {`pale_skin`, `narrow_eyes`}. We represent these logical operators as binary classifiers and train them on CelebA-STS. We get the labels for these logical operators by applying the same logical operation over labels for the simple and the complex attribute. We then test the resiliency of these logical operators by shifting the distribution of the simple attribute and the complex attribute, one at a time. Specifically, we vary the proportion (in percentages) of images with simple attribute (or complex attribute) in test set while keeping the proportion of images with complex attribute (or simple attribute) fixed. We show the results in Figure 1 and Figure 4. We observe that the success rate of the logical operators vary less when we shift the distribution of the simple attributes

$\{\texttt{bald}, \texttt{big\_lips}\}$.

Implementation details We parameterize multi-head binary classifier, predicting presence of 40 different facial attribute, with MobileNet (Howard et al., 2017). We train it with Adam optimizer (Kingma and Ba, 2014) for 100 epochs using a learning rate of 0.001 and a batch size of 64. We use the same architecture and the training hyperparameters for learning logical operators f_{OR} and f_{XOR}. We borrow these hyperparameters and the code for running our experiments from the github repo https://github.com/suikei-wang/Facial-Attributes-Classification.

F.5. Imitation learning on Robotic pusher arm environment

Environment Description We use Robotic pusher arm environment adapted from Ajay et al. (2022); Gupta et al. (2018) where the goal is to push the red cube to the green circle. When the red cube reaches the green circle, the agent gets a reward of +1. The state space is 12-dimensional consisting of mass of red cube (1), dampness parameter for each joint (1), joint angles (3) and velocities (3) of the gripper, COM of the gripper (2) and position of the red cube (2). The green circle's position is fixed and at an initial distance of 0.5 from COM of the gripper. The red cube (of size 0.03) is initially at a distance of 0.1 from COM of the gripper and at an angle $\pi/4$. During training, at beginning for every episode, we sample $m \sim \mathcal{N}(60, 15)$ and $d \sim \mathcal{N}(0.5, 0.1)$. The task horizon is 60 timesteps.

Expert Policy To obtain expert policy that provides data for imitation learning and for training dynamics models $(p_{\phi_i}(s_{t+1}|s_t, a_t, i),\ i \in \{d, m\})$, we train a policy $\pi_{\text{exp}}(a|s, m, d)$ with Soft-Actor-Critic (Haarnoja et al., 2018) for $10e6$ environment steps.

Determining *simple* factor To determine which of the two is "simpler", we measure generalization error on the auxillary task of predicting next-step dynamics where one of $\{m, d\}$ is held fixed, and the other drawn from a certain distribution, fixed across both testing and training. We learn two dynamics model $p_{\phi_m}(s_{t+1}|s_t, a_t, m)$ and $p_{\phi_d}(s_{t+1}|s_t, a_t, d)$ on two separate datasets $\mathcal{D}^{train}_{\text{dyn},m}$ and $\mathcal{D}^{train}_{\text{dyn},d}$. Both $\mathcal{D}^{train}_{\text{dyn},m}$ and $\mathcal{D}^{train}_{\text{dyn},d}$ contain 100 expert trajectories each, with varying m and d respectively while keeping the other factor fixed. While $m \sim \mathcal{N}(60, 15)$ and $d = 0.5$ in $\mathcal{D}^{train}_{\text{dyn},m}$, $d \sim \mathcal{N}(0.5, 0.1)$ and $m = 60$ in $\mathcal{D}^{train}_{\text{dyn},d}$. To evaluate learned dynamics models, we generate $\mathcal{D}^{test}_{\text{dyn},m}$ and $\mathcal{D}^{test}_{\text{dyn},d}$ in same way as their training counterparts. On training datasets, we find mean squared error $(\frac{1}{|\mathcal{D}|}\|s_{t+1} - p_{\phi_i}(s_{t+1}|s_t, a_t, i)\|_2^2,\ i \in \{m, d\})$ of $p_{\phi_m}(s_{t+1}|s_t, a_t, m)$ and $p_{\phi_d}(s_{t+1}|s_t, a_t, d)$ to be 0.062 and 0.183 respectively. On test datasets, we find the mean squared error of $p_{\phi_m}(s_{t+1}|s_t, a_t, m)$ and $p_{\phi_d}(s_{t+1}|s_t, a_t, d)$ to be 0.081 and 0.237 respectively. Since $p_{\phi_m}(s_{t+1}|s_t, a_t, m)$ has smaller generalization error, we consider object `mass` as simple. Intuitively, object `mass` only affects the dynamics of the system when the robotic arm is in contact with the object. In contrast, the joints' `dampness` affects the way the robotic arm moves and hence affects the system's dynamics independent of whether the robotic arm is in contact with the object.

Implementation details We parameterize dynamics model $(p_{\phi_i}(s_{t+1}|s_t, a_t, i),\ i \in \{d, m\})$, expert policy $\pi_{\text{exp}}(a|s, m, d)$ and imitator policy $\pi_\theta(a|s, m, d)$ with a 3-layered Multi-layer perception (MLP) having hidden dimension of 512 and ReLU activation. We train dynamics model and imitator policy with Adam optimizer (Kingma and Ba, 2014) for 100 epochs using a learning rate of 0.001 and a batch size of 128.

Attribute	Train Accuracy	Test Accuracy
5_o_Clock_Shadow	0.952 ± 0.003	0.945 ± 0.008
Arched_Eyebrows	0.879 ± 0.01	0.84 ± 0.006
Attractive	0.846 ± 0.001	0.828 ± 0.003
Bags_Under_Eyes	0.872 ± 0.004	0.845 ± 0.006
Bald	$\mathbf{0.991 \pm 0.002}$	$\mathbf{0.988 \pm 0.004}$
Bangs	0.968 ± 0.006	0.96 ± 0.005
Big_Lips	$\mathbf{0.8 \pm 0.008}$	$\mathbf{0.716 \pm 0.009}$
Big_Nose	0.867 ± 0.002	0.839 ± 0.004
Black_Hair	0.918 ± 0.007	0.896 ± 0.006
Blond_Hair	0.963 ± 0.005	0.957 ± 0.003
Blurry	0.966 ± 0.008	0.961 ± 0.005
Brown_Hair	0.886 ± 0.009	0.885 ± 0.009
Bushy_Eyebrows	0.929 ± 0.007	0.922 ± 0.002
Chubby	0.964 ± 0.003	0.948 ± 0.005
Double_Chin	0.971 ± 0.002	0.961 ± 0.006
Eyeglasses	0.998 ± 0.003	0.996 ± 0.006
Goatee	0.978 ± 0.002	0.974 ± 0.003
Gray_Hair	0.984 ± 0.005	0.98 ± 0.002
Heavy_Makeup	0.939 ± 0.007	0.918 ± 0.005
High_Cheekbones	0.898 ± 0.006	0.876 ± 0.004
Male	0.989 ± 0.001	0.979 ± 0.003
Mouth_Slightly_Open	0.957 ± 0.004	0.936 ± 0.008
Mustache	0.975 ± 0.005	0.97 ± 0.007
Narrow_Eyes	$\mathbf{0.915 \pm 0.002}$	$\mathbf{0.875 \pm 0.004}$
No_Beard	0.971 ± 0.009	0.96 ± 0.007
Oval_Face	0.795 ± 0.002	0.758 ± 0.005
Pale_Skin	$\mathbf{0.97 \pm 0.008}$	$\mathbf{0.967 \pm 0.005}$
Pointy_Nose	0.795 ± 0.011	0.774 ± 0.009
Receding_Hairline	0.952 ± 0.003	0.939 ± 0.008
Rosy_Cheeks	0.959 ± 0.004	0.95 ± 0.009
Sideburns	0.981 ± 0.002	0.978 ± 0.003
Smiling	0.948 ± 0.006	0.928 ± 0.005
Straight_Hair	0.856 ± 0.003	0.831 ± 0.009
Wavy_Hair	0.87 ± 0.003	0.833 ± 0.004
Wearing_Earrings	0.923 ± 0.004	0.9 ± 0.002
Wearing_Hat	0.994 ± 0.009	0.989 ± 0.005
Wearing_Lipstick	0.946 ± 0.003	0.934 ± 0.007
Wearing_Necklace	0.895 ± 0.004	0.87 ± 0.006
Wearing_Necktie	0.97 ± 0.001	0.965 ± 0.002
Young	0.912 ± 0.002	0.877 ± 0.004

Table 1. Train and test accuracy of multi-head binary classifier for each attribute in CelebA dataset. We consider attributes with low generalization error as *simple* and attributes with high generalization error as *complex*. We highlight the selected *simple* (Bald, Pale Skin) and *complex* (Big Lips, Narrow Eyes) features. We report mean and standard error across 4 replicates.