# A Statistical Perspective on Retrieval-Based Models

**Soumya Basu** [*1] **Ankit Singh Rawat** [*2] **Manzil Zaheer** [*3]

## Abstract

Many modern high-performing machine learning models increasingly rely on scaling up models, e.g., transformer networks. Simultaneously, a parallel line of work aims to improve the model performance by augmenting an input instance with other (labeled) instances during inference. Examples of such augmentations include task-specific prompts and similar examples retrieved from the training data by a nonparametric component. Despite a growing literature showcasing the promise of these *retrieval-based models*, their theoretical underpinnings remain under-explored. In this paper, we present a formal treatment of retrieval-based models to characterize their performance via a novel statistical perspective. In particular, we study two broad classes of retrieval-based classification approaches: First, we analyze a *local learning* framework that employs an *explicit* local empirical risk minimization based on retrieved examples for each input instance. Interestingly, we show that breaking down the underlying learning task into local sub-tasks enables the model to employ a *low complexity* parametric component to ensure good overall performance. The second class of retrieval-based approaches we explore learns a global model using kernel methods to directly map an input instance and retrieved examples to a prediction, without explicitly solving a local learning task.

## 1. Introduction

As our world is complex, we need expressive machine learning (ML) models to make high-accuracy predictions on real-world problems. There are multiple ways to increase the expressiveness of an ML model. A popular way is to homogeneously scale the size of a parametric model, such as neural networks, which has been behind many recent high-performance models such as GPT-3 (Brown et al., 2020) and ViT (Dosovitskiy et al., 2021). Their performance (accuracy) exhibits a monotonic behavior with increasing model size, as demonstrated by "scaling laws" (Kaplan et al., 2020; Hoffmann et al., 2022). Such large models, however, have their own limitations, including high computation cost, catastrophic forgeting (hard to adapt to changing data), lack of provenance, and poor explainability. Classical instance-based models (Fix & Hodges, 1989), on the other hand, offer many desirable properties by design — efficient data structures, incremental learning (easy addition and deletion of knowledge), and some provenance for its prediction based on the nearest neighbors w.r.t. the input. However, these models often suffer from weaker empirical performance as compared to deep parametric models.

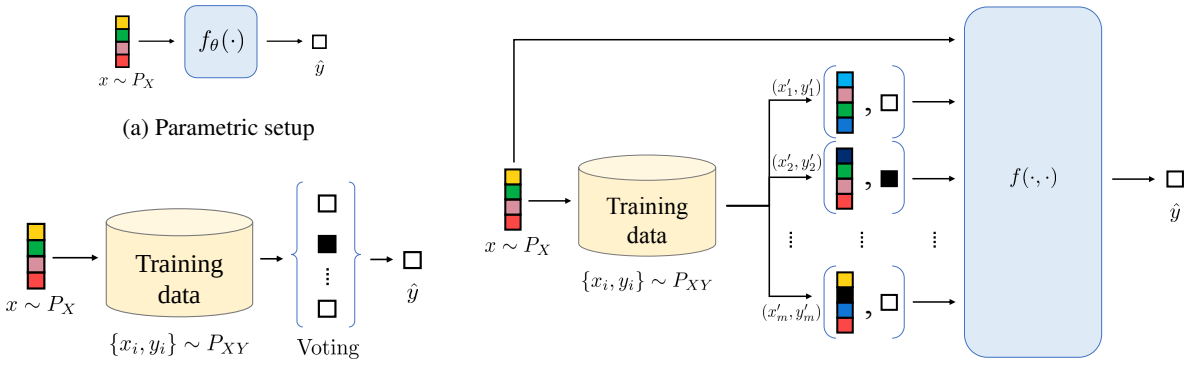
Increasingly, a middle ground combining the two paradigms and retaining the best of both worlds is becoming popular across various domains, ranging from natural language (Das et al., 2021; Wang et al., 2022; Liu et al., 2022; Izacard et al., 2022), to vision (Liu et al., 2015; 2019; Iscen et al., 2022; Long et al., 2022), to reinforcement learning (Blundell et al., 2016; Pritzel et al., 2017; Ritter et al., 2020), to even protein structure prediction (Cramer, 2021). In such approaches, given a test input, one first retrieves relevant entries from a data index and then processes the retrieved entries along with the test input to make the final predictions using an ML model. This process is visualized in Fig. 1c.

While classical learning setups (cf. Fig. 1a and 1b) have been studied extensively over decades, even basic properties and trade-offs pertaining to retrieval-based models (cf. Fig. 1c), despite their aforementioned remarkable successes, remain highly under-explored. Most of the existing efforts on retrieval-based models solely focus on developing end-to-end *domain-specific* models, without identifying the key dataset properties or structures that are critical in realizing performance gains by such models. Furthermore, at first glance, due to the highly dependent nature of an input and the associated retrieved set, direct application of existing statistical learning techniques does not appear as straightforward. This prompts a natural question:

*What is the right theoretical framework to rigorously showcase the value of the retrieved set in ensuring superior performance of modern retrieval-based models?*

[*]Equal contribution; in alphabetical order [1]Google, Mountain View, USA [2]Google Research, New York, USA [3]Google DeepMind, New York, USA. Correspondence to: Soumya Basu <basusoumya@google.com>.

*Proceedings of the $40^{th}$ International Conference on Machine Learning*, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).




(a) Parametric setup

(b) Instance-based learning setup

(c) Modern retrieval-based setup

*Figure 1.* An illustration of a retrieval-based classification model. Given an input instance $x$, similar to an instance-based model, it retrieves similar (labeled) examples $\mathcal{R}^x = \{(x'_j, y'_j)\}_j$ from training data. Subsequently, it processes input instance along with the retrieved examples (potentially via a nonparametric method) to make the final prediction $\hat{y} = f(x, \mathcal{R}^x)$.

In this paper, we take the first step towards answering this question, while focusing on the classification setting (Sec. 2.1). We begin with the hypothesis that the model might be using the retrieved set to do *local learning* implicitly and then adapt its predictions to the neighborhood of the test point. Multiple recent works (Garg et al., 2022; Akyürek et al., 2022; von Oswald et al., 2022) have studied the feasibility of such a mechanism in widely popular Transformer models. Notably, these works show that a Transformer network can emulate gradient descent to optimize a local learning objective when presented with multiple labeled examples as inputs. Such local learning is potentially beneficial in cases where the underlying task has a local structure, where a much simpler function class suffices to explain the data in a given local neighborhood but overall the data can be complex (formally defined in Sec. 2.2). For example, to solve an issue at hand (a problem instance), it is often faster to search for solutions to similar problems on Stackoverflow and utilize those (i.e., locally learning from the retrieved similar labeled examples) than understanding the whole system (i.e., learning the entire global function).

Inspired by Bottou & Vapnik (1992), we analyze an *explicit* local learning framework: For each test input, 1) we retrieve a few (labeled) training examples located in the vicinity of the test input, 2) train a local model by performing empirical risk minimization (ERM) with only these retrieved examples – *local ERM*; and 3) apply the resulting local model to make prediction on the test input. For the aforementioned retrieval-based local ERM, we derive finite sample generalization bounds that highlight a trade-off between the complexity of the underlying function class and size of the neighborhood where local structure of the data distribution holds in Sec. 3. Under this assumption of local regularity, we show that by using a much simpler function class for the local model, we can achieve a similar loss/error to that of a complex global model (Thm. 3.7). Thus, we show that breaking down the underlying learning task into local sub-tasks enables the model to employ a *low complexity* parametric component to ensure good global accuracy via a retrieval-based model.

We acknowledge that such local learning cannot be the complete picture behind the effectiveness of retrieval-based models. As noted in Zakai & Ritov (2008), there always exists a model with global component that is more "preferable" to a local-only model. In Sec. 3.4, we extend local ERM to a two-stage setup: First learn a global representation using entire dateset, and then utilize the representation at the test time while solving the local ERM as previously defined. This enables the local learning to benefit from good quality global representations, especially in sparse data regions.

Finally, we move beyond explicit local learning to a setting that resembles more closely the empirically successful systems such as REINA (Wang et al., 2022), WebGPT (Nakano et al., 2021), and AlphaFold (Cramer, 2021): A model that directly learns to predict from the input instance and associated retrieved similar examples end-to-end. Towards this, we take a preliminary step in Sec. 4 by studying a novel formulation of classification over an extended feature space (to account for the retrieved examples) by using kernel methods (Deshmukh et al., 2019).

To summarize, our main contributions include: 1) Setting up a formal framework for classification via retrieval-based models under local structure; 2) Finite sample analysis of explicit local learning framework; 3) Comparison with simple parametric and nonparametric paradigms 4) Extending the analysis to a globally learnt model; and 5) Providing the first rigorous treatment of an end-to-end retrieval-based model to study its generalization by using kernel-based learning.

## 2. Problem setup

We first provide a brief background on (multiclass) classification along with the necessary notations. Subsequently, we discuss the problem setup considered in this paper, which

deals with designing retrieval-based classification models for the data distributions with local structures.

### 2.1. Multiclass classification

In this work, we restrict ourselves to (multi-class) classification setting, with access to $n$ training examples $\mathcal{S} = \{(x_i, y_i)\}_{i\in[n]} \subset \mathcal{X} \times \mathcal{Y}$, sampled i.i.d. from the data distribution $\mathsf{D} := \mathsf{D}_{X,Y}$. Let $\mathbb{P}_{\mathsf{D}}(A) := \mathbb{E}_{(X,Y)\sim\mathsf{D}}\left[\mathbf{1}_{\{A\}}\right]$ for any random variable $A$.

Given $\mathcal{S}$, one is interested in learning a classifier $h : \mathcal{X} \to \mathcal{Y}$ that minimizes miss-classification error. It is common to define a classifier via a scorer $f : x \mapsto \big(f_1(x), \ldots, f_{|\mathcal{Y}|}(x)\big) \in \mathbb{R}^{|\mathcal{Y}|}$ that assigns a score to each class in $\mathcal{Y}$ for an instance $x$. For a scorer $f$, the corresponding classifier takes the form:

$$h_f(x) = \arg\max_{y\in\mathcal{Y}} f_y(x).$$

Given a set of scorers $\mathcal{F}^{\text{global}} \subseteq \{f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}\}$, learning a model implies finding a scorer in $\mathcal{F}^{\text{global}}$ that minimizes the miss-classification error or expected $0/1$-loss:

$$f^*_{0/1} = \arg\min_{f\in\mathcal{F}^{\text{global}}} \mathbb{P}_{\mathsf{D}}(h_f(X) \neq Y). \tag{1}$$

One typically employs a surrogate loss (Bartlett et al., 2006) $\ell$ for the miss-classification error $\mathbb{1}_{\{h_f(X)\neq Y\}}$ and aims to minimize the associated population risk:

$$R_\ell(f) = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\big[\ell\big(f(X), Y\big)\big].$$

Since the underlying data distribution $\mathsf{D}$ is only accessible via examples in $\mathcal{S}$, one learns a good scorer by minimizing the (global) *empirical* risk over the function class $\mathcal{F}^{\text{global}}$ as follows:

$$\hat{f} = \arg\min_{f\in\mathcal{F}^{\text{global}}} \frac{1}{n}\sum_{i\in[n]} \ell\big(f(x_i), y_i\big). \tag{2}$$

We denote $\widehat{R}_\ell(f) := \frac{1}{n}\sum_{i\in[n]} \ell\big(f(x_i), y_i\big)$.

### 2.2. Classification with local structure

In this work, we assume that the underlying data distribution $\mathsf{D}$ has a local structure, where a much simpler (parametric) function class suffices to explain the data in each local neighborhood. Formally, for $x \in \mathcal{X}$ and $r > 0$, we define $\mathcal{B}^{x,r} := \{x' \in \mathcal{X} : \mathbb{d}(x, x') \leq r\}$, an $r$-radius ball around $x$, w.r.t. a metric $\mathbb{d} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $\mathsf{D}^{x,r}$ be the data distribution restricted to $\mathcal{B}^{x,r}$, i.e.,

$$\mathsf{D}^{x,r}(A) = \mathsf{D}(A)/\mathsf{D}\left(\mathcal{B}^{x,r} \times \mathcal{Y}\right) \quad A \subseteq \mathcal{B}^{x,r} \times \mathcal{Y}. \tag{3}$$

Further, let us define the local population risk of a function $f$ at a given instance $x \in \mathcal{X}$:

$$R^x_\ell(f) = \mathbb{E}_{(X',Y')\sim\mathsf{D}^{x,r}}\big[\ell\big(f(X'), Y'\big)\big].$$

Now, the *local structure condition* of the data distribution ensures that, for each $x \in \mathcal{X}$, there exists a low-complexity function class $\mathcal{F}^x$, with $|\mathcal{F}^x| \ll |\mathcal{F}^{\text{global}}|$, that approximates the Bayes optimal (w.r.t. $\mathcal{F}^{\text{global}}$) for the *local classification problem* defined by $\mathsf{D}^{x,r}$. That is, for a given $\varepsilon_{\mathcal{X}} > 0$ and $\forall\, x \in \mathcal{X}$, we have that [1]

$$\min_{f\in\mathcal{F}^x} R^x_\ell(f) \leq \min_{f\in\mathcal{F}^{\text{global}}} R^x_\ell(f) + \varepsilon_{\mathcal{X}}. \tag{4}$$

As an example, if $\mathcal{F}^{\text{global}}$ is linear in $\mathbb{R}^d$ (possibly dense) with bounded norm $\tau$, then $\mathcal{F}^x$ can be a simpler function class such as linear in $\mathbb{R}^d$ with sparsity $k \ll d$ and with bounded norm $\tau_x \leq \tau$.

### 2.3. Retrieval-based classification model

This work focuses on retrieval-based methods that can leverage the aforementioned local structure of the data distribution. In particular, we focus on two such approaches:

**Local empirical risk minimization.** Given a (test) instance $x$, the *local* empirical risk minimization (ERM) approach first retrieves a neighboring set $\mathcal{R}^x = \{(x'_j, y'_j)\} \subseteq \mathcal{S}$. Subsequently, it identifies a (local) scorer $\hat{f}^x$ from a 'simple' function class $\mathcal{F}^{\text{loc}} \subset \{f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}\}$ as follows:

$$\hat{f}^x = \arg\min_{f\in\mathcal{F}^{\text{loc}}} \frac{1}{|\mathcal{R}^x|}\sum_{(x',y')\in\mathcal{R}^x} \ell\big(f(x'), y'\big). \tag{5}$$

By convention if $|\mathcal{R}^x| = 0$, $\hat{f}^x \in \mathcal{F}^{\text{loc}}$ is chosen arbitrarily.

Note that the local ERM approach requires solving a local learning task for each test instance. Such a local learning algorithms was introduced by Bottou & Vapnik (1992). Another point worth mentioning here is that (5) employs the same function class $\mathcal{F}^{\text{loc}}$ for each $x$, whereas the local structure assumption (cf. (4)) allows for an instance dependent function class $\mathcal{F}^x$. We consider $\mathcal{F}^{\text{loc}}$ that approximates $\cup_{x\in\mathcal{X}}\mathcal{F}^x$ closely. In particular, we assume that, for some $\varepsilon_{\text{loc}} > 0$, we have $\forall\, x \in \mathcal{X}$ that

$$\min_{f\in\mathcal{F}^{\text{loc}}} R^x_\ell(f) \leq \min_{f\in\mathcal{F}^x} R^x_\ell(f) + \varepsilon_{\text{loc}}. \tag{6}$$

Continuing with the example following (4), where $\mathcal{F}^x$ is linear with sparsity $k \ll d$ and bounded norm $\tau_x$, one can take $\mathcal{F}^{\text{loc}}$ to be linear with the same sparsity $k$ and bounded norm $\tau' < \sup_{x\in\mathcal{X}} \tau_x$.

**Classification with extended feature space.** We also consider the setting where the scorer $f$ can implicitly solve the local-ERM using retrieved neighboring labeled instances to make the classification prediction. In other words, the scorer directly maps the augmented input $x \times \mathcal{R}^x \in \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^\star$ to per-class scores. One can learn such a scorer over *extended* feature space $\mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^\star$ as follows:

$$\hat{f}^{\text{ex}} = \arg\min_{f\in\mathcal{F}^{\text{ex}}} \hat{R}^{\text{ex}}_\ell(f), \tag{7}$$

where $\hat{R}^{\text{ex}}_\ell(f) := \frac{1}{n}\sum_{i\in[n]} \ell\big(f\big(x_i, \mathcal{R}^{x_i}\big), y_i\big)$ and a function class of interest over the extended space is denoted

[1] As stated, we require the local structure condition to hold for each $x$. This can be relaxed to hold with high probability with the increased complexity of exposition.

as $\mathcal{F}^{\text{ex}} \subset \{f : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^\star \to \mathbb{R}^{|\mathcal{Y}|}\}$. Examples of such a function class include prompting transformers with the retrieved labeled examples. Moreover, it has been recently shown that a transformer can express certain algorithms for optimizing a local learning objective based on the examples from the prompt using gradient descent (Garg et al., 2022; Akyürek et al., 2022; von Oswald et al., 2022).

Our goal is to develop a statistical understanding of these two retrieval-based methods for classification when the underlying data distribution has local structure. We present our theoretical treatment of local ERM and classification with extended feature space in Sec. 3 and 4, respectively.

## 3. Local empirical risk minimization

In this section, our objective is to characterize the excess risk of local ERM. In particular, we aim to bound

$$\mathbb{E}_{(X,Y)\sim \mathsf{D}}\big[\ell(\hat{f}^X(X), Y) - \ell(f^*(X), Y)\big]. \tag{8}$$

Note that $\hat{f}^X$ (cf. (5)) in the above equation is a function of $\mathcal{R}^X$, and expectation over $\mathcal{R}^X$ is taken implicitly.

### 3.1. Assumptions

Before presenting an excess risk bound for the local ERM method, we introduce various necessary definitions and assumptions that play a critical role in our analysis.

We define the *margin* of scorer $f$ at a given label $y \in \mathcal{Y}$ as

$$\gamma_f(x, y) = f_y(x) - \max_{y' \neq y} f_{y'}(x). \tag{9}$$

In order to ensure the margin of the scorer $f$ has smooth deviation as $x$ varies, we introduce $L$-coordinate Lipschitz condition: A scorer $f$ is *L-coordinate Lipschitz* iff for all $y \in \mathcal{Y}$ and $x, x' \in \mathcal{X}$, we have

$$|f_y(x) - f_y(x')| \leq L\|x - x'\|_2. \tag{10}$$

Following Döring et al. (2018), we define the *weak margin condition* for a scorer $f$: Given a distribution $\mathsf{D}$, a scorer $f$ satisfies $(\alpha, c)$-weak margin condition iff, for all $t \geq 0$,

$$\mathbb{P}_{(X,Y)\sim \mathsf{D}}(|\gamma_f(X, Y)| \leq t) \leq c\, t^\alpha. \tag{11}$$

One of the key assumptions that we rely on is the existence of an underlying scorer $f^{\text{true}}$ that explains the true labels, while ensuring the weak margin condition. Here, we note that the true function $f^{\text{true}}$ may neither lie in the function class $\mathcal{F}^{\text{global}}$, nor in $\mathcal{F}^{\text{loc}}$.

**Assumption 3.1** (True scorer function)**.** *There exists a scorer $f^{\text{true}}$ such that, for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $f^{\text{true}}$ generates the true label, i.e., $\gamma_{f^{\text{true}}}(x, y) > 0$. Furthermore, we assume $f^{\text{true}}$ is $L_{\text{true}}$-coordinate Lipschitz, and satisfies the $(\alpha_{\text{true}}, c_{\text{true}})$-weak margin condition.*

Furthermore, we restrict ourselves to smooth loss functions that act on the margin of a scorer (cf. (9)).

**Assumption 3.2** (Margin-based Lipschitz loss)**.** *For any given example $(x, y)$ and any scorer $f$, we have $\ell(f(x), y) = \ell(\gamma_f(x, y))$ and $\ell$ is a decreasing function of the margin. Furthermore, the loss function $\ell$ is $L_\ell$-Lipschitz function, i.e., $|\ell(\gamma) - \ell(\gamma')| \leq L_\ell|\gamma - \gamma'|$, $\forall \gamma \geq \gamma'$.*

Recall that $\mathcal{R}^x$ corresponds to the samples in $\mathcal{S}$ that belong to $\mathcal{B}^{x,r}$; hence, it follows the distribution $\mathsf{D}^{x,r}$. For the rest of the paper, we limit ourselves to $\mathcal{X} \subseteq \mathbb{R}^d$. We can extend this to more general metric spaces with the increased complexity of exposition. Let the density of the distribution of $x \in \mathcal{X} \subseteq \mathbb{R}^d$ be $\rho_{\mathsf{D}}(x)$. A common assumption in the nonparametric estimation literature is the weak density condition (see, e.g., Döring et al., 2018). Moreover, we need to ensure that with high probability the density $\rho_{\mathsf{D}}(x)$ is not too low. We do so following the idea of density level sets from Steinwart (2011). Accordingly, we make the following assumption.

**Assumption 3.3** (Data regularity condition)**.**

1. (Weak density condition) *There exists constants $c_{\text{wdc}} > 0$, and $\delta_{\text{wdc}} > 0$, such that for all $x \in \mathcal{X}$ and $\rho_{\mathsf{D}}(x)r^d \leq \delta^d_{\text{wdc}}$,*

$$\mathbb{P}_{X'\sim \mathsf{D}}[\mathfrak{d}(X', x) \leq r] \geq c^d_{\text{wdc}}\rho_{\mathsf{D}}(x)r^d.$$

2. (Density level-set) *There exists a function $f_\rho(\delta)$ with $f_\rho(\delta) \to 0$ as $\delta \to 0$, such that for any $\delta > 0$,*

$$\mathbb{P}_{X\sim \mathsf{D}}[\rho_{\mathsf{D}}(X) \leq f_\rho(\delta)] \leq \delta. \tag{12}$$

For example, for $d$-dimensional multivariate Gaussian with the covariance matrix $\Sigma$, we have $f_\rho(\delta) = \Theta(2^{-d/2}|\Sigma|^{-1/2}\delta \ln(1/\delta)^{-d/2})$. This result can be extended to mixture of Gaussian and sub-gaussian random variables (see Appendix B.6 for details).

**Assumption 3.4** (Weak+ density condition)**.** *There exists constants $c_{\text{wdc}+} \geq 0$, and $\alpha_{\text{wdc}+} > 0$, such that for all $x \in \mathcal{X}$ and $r \in [0, r_{\max}]$,*

$$\Big|\frac{\mathbb{P}_{X'\sim \mathsf{D}}[\mathfrak{d}(X', x) \leq r]}{\rho_{\mathsf{D}}(x) vol_d(r)} - 1\Big| \leq c_{\text{wdc}+} r^{\alpha_{\text{wdc}+}}.$$

The above assumption implies Assumption 3.3.1. We will show that under Assumption 3.4 the local ERM error bounds can be tightened further. For example, in $d$-dimensional multivariate Gaussian with the covariance matrix $\Sigma$, we have $c_{\text{wdc}+} = \frac{d\lambda_{\max}(\Sigma^{-1})}{2(d+2)}$, and $\alpha_{\text{wdc}+} = 2$ for $r_{\max} = \sqrt{(d+2)\lambda_{\max}(\Sigma)}$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue.

### 3.2. Excess risk bound for local ERM

We now proceed to our main results on the excess risk bound of local ERM. Recall that, at $x \in \mathcal{X}$, $f^{x,*}$ denotes the minimizer of the population version of the local loss, and $f^*$ the population risk minimizer for the global loss, i.e.,

$$f^{x,*} = \operatorname*{arg\,min}_{f\in\mathcal{F}^{\text{loc}}} \mathcal{R}^x_\ell(f); \quad f^* = \operatorname*{arg\,min}_{f\in\mathcal{F}^{\text{global}}} \mathcal{R}_\ell(f). \tag{13}$$

To bound the excess risk defined in Eq. (8), we first obtain the following upper bound on (8).

**Lemma 3.5** (Risk decomposition)**.** *The expected excess risk of the local ERM solution $\hat{f}^X$ is bounded as*

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\ell(\hat{f}^X(X),Y)-\ell(f^*(X),Y)\Big]\\
&\le \underbrace{\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[R^X_\ell(f^{X,*})-R^X_\ell(f^*)\Big]}_{\text{Local vs Global Population Optimal Risk}}\\
&+\underbrace{\sum_{\mathcal{F}\in\{\mathcal{F}^{\text{global}},\mathcal{F}^{\text{loc}}\}}\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}}\big|R^X_\ell(f)-\ell(f(X),Y)\big|\Big]}_{\text{Global and Local: Sample vs Retrieved Set Risk}}\\
&+\underbrace{\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}}\Big|R^X_\ell(f)-\hat{R}^X_\ell(f)\Big|\Big]}_{\text{Generalization of Local ERM}}\\
&+\underbrace{\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\Big|R^X_\ell(f^{X,*})-\hat{R}^X_\ell(f^{X,*})\Big|\Big]}_{\text{Central Absolute Moment of } f^{X,*}}.
\end{aligned}$$

We delegate the proof of Lem. 3.5 to Appendix B. Now, as a strategy to obtain desired excess risk bounds, we separately bound the four terms appearing in Lem. 3.5. Note that the first term captures the expected difference between the loss incurred by global population optima $f^* \in \mathcal{F}^{\text{global}}$ and the local population optima $f^{x,*} \in \mathcal{F}^{\text{loc}}$ in a local region around the test instance $x$. The second term aims to capture the loss for a scorer evaluated at $x$ vs. the expected value of the loss for the scorer at a random instance sampled in the local region of $x$ based on $\mathsf{D}^{x,r}$. The third term corresponds to the standard 'generalization error' for the local ERM with respect to the local data distribution $\mathsf{D}^{X,r}$, whereas the fourth term is the empirical variation of the local population optima $f^{X,*}$ around its population mean under $\mathsf{D}^{X,r}$.

Let the coordinate-Lipschitz constants for scorers in $\mathcal{F}^{\text{loc}}$ and $\mathcal{F}^{\text{global}}$ be $L_{\text{loc}}$ and $L_{\text{global}}$, respectively. We define a function class $\mathcal{G}(X,Y) := \{(x',y') \mapsto \ell(\gamma_f(\cdot,\cdot)) - \ell(\gamma_f(X,Y)) : f \in \mathcal{F}^{\text{loc}}\}$. Here, by subtracting $\ell\big(f(X),Y\big)$ from the loss, we center the losses on $\mathcal{R}^X$ for any function $f \in \mathcal{F}^{\text{loc}}$, and obtain a tighter bound by utilizing the local structure of the distribution $\mathsf{D}^{X,r}$. For any $L > 0$, for notational convenience let us define

$$\begin{aligned}
&\mathcal{M}_r(L;\ell,f_{\text{true}},\mathcal{F}) :=\\
&\quad 2L_\ell\Big(Lr+\big(2\|\mathcal{F}\|_\infty - Lr\big)c_{\text{true}}\big(2L_{\text{true}}r\big)^{\alpha_{\text{true}}}\Big).
\end{aligned} \tag{14}$$

For any $x \in \mathcal{X}$, the weak density condition provides high probability lower bound on the size of the retrieved set $\mathcal{R}^x$.

**Proposition 3.6.** *Under the Assumption 3.3, for any $x \in \mathcal{X}$, radius $r > 0$, and $\delta > 0$,*

$$\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^x| < N(r,\delta)\big] \le \delta, \tag{15}$$

*for* $N(r,\delta) = n\big(c^d_{\text{wdc}}\min\{f_\rho(\delta/2)r^d, \delta^d_{\text{wdc}}\} - \sqrt{\frac{\log(2/\delta)}{2n}}\big)$.

Now, by controlling different terms appearing in the bound in Lem. 3.5, we obtain the following.

**Theorem 3.7** (**Excess risk bound**)**.** *Let* (4) *and* (6)*; and Assumptions 3.1, 3.2 and 3.3 hold. For any $\delta > 0$, and $N(r,\delta)$ as defined in Proposition 3.6, the expected excess risk of the local ERM solution $\hat{f}^X$ is bounded as*

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\ell(\hat{f}^X(X),Y)-\ell(f^*(X),Y)\Big]\\
&\le \underbrace{(\varepsilon_{\mathcal{X}}+\varepsilon_{\text{loc}})}_{\text{Local vs Global Optimal loss (I)}}\\
&+\underbrace{\mathcal{M}_r(L_{\text{loc}};\ell,f_{\text{true}},\mathcal{F}^{\text{loc}})+\mathcal{M}_r(L_{\text{global}};\ell,f_{\text{true}},\mathcal{F}^{\text{global}})}_{\text{Global and Local: Sample vs Retrieved Set Risk (II)}}\\
&+\underbrace{\begin{boxed}\begin{aligned}&\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)\big|\mathcal{R}^X \ge N(r,\delta)\Big]\\&+5\mathcal{M}_r(L_{\text{loc}};\ell,f_{\text{true}},\mathcal{F}^{\text{loc}})\sqrt{\frac{2\ln(4/\delta)}{N(r,\delta)}}\\&+8\delta L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty,\end{aligned}\end{boxed}}_{\text{Generalization of Local ERM and Central Absolute Moment of } f^{X,*}\ \text{(III)}}
\end{aligned}$$

*where $\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)$ denotes the empirical Rademacher complexity of $\mathcal{G}(X,Y)$. Under Assumption 3.4 and $r \le r_{\max}$, Sample vs Retrieved Set Risk* (II) *is $O(c_{wdc+}r^{\alpha_{wdc+}})$.*

The above result shows a trade-off in *approximation* vs. *generalization* error as retrieval radius $r$ varies.

**Approximation error.** It comprises two components, defined by (I) and (II) in Thm. 3.7. $\varepsilon_{\mathcal{X}}$ shows the gap in approximating the $r$-radius neighborhood around $X$ with a simple local function class $\mathcal{F}^X$ which varies with $X \in \mathcal{X}$. Next, $\varepsilon_{\text{loc}}$ shows the gap in approximating the union of the local function class $\cup_{x\in\mathcal{X}}\mathcal{F}^x$ with a single function class $\mathcal{F}^{\text{loc}}$ (possibly with smaller complexity) but while allowing for choosing a different optimizer $f^X \in \mathcal{F}^{\text{loc}}$ for each $X \in \mathcal{X}$. Both $\varepsilon_{\mathcal{X}}$ and $\varepsilon_{\text{loc}}$ typically increases with $r$.

The second component of the approximation error (II) corresponds to the difference of risk for the sample $X$ and the retrieved set $\mathcal{R}^X$ for $\mathcal{F}^{\text{global}}$ and $\mathcal{F}^{\text{loc}}$, i.e., $\mathcal{M}_r(L_{\text{global}};\ell,f_{\text{true}},\mathcal{F}^{\text{global}})$ and $\mathcal{M}_r(L_{\text{loc}};\ell,f_{\text{true}},\mathcal{F}^{\text{loc}})$. Eq. (14) suggests that the terms increase as $O(\text{poly}(r))$. When the data follows multivaraite Gaussian then term $(II)$ increases as $O(r^2)$.

**Generalization error.** It (III) depends on the size of the retrieved set $\mathcal{R}^X$ and the Rademacher complexity of $\mathcal{G}(X,Y)$ which is induced by $\mathcal{F}^{\text{loc}}$. With increasing radius $r$, the term $N(r,\delta)$ increases. The Rademacher complexity decays with increasing radius, $r$, typically at the rate of $O(1/\sqrt{N(r,\delta)})$. Thus, under the local ERM setting the total approximation error increases with increasing radius $r$, given $\mathcal{F}^{\text{loc}}$ is fixed. On the contrary, the generalization error decreases with increasing radius $r$ for a fixed $\mathcal{F}^{\text{loc}}$. This suggests a trade-off between the approximation and generalization error as we make a design choice about $r$. (We empirically validate this in Fig. 3.) Due to centering within the set $\mathcal{G}(X,Y)$ we have

the upper bound on this term as $\mathcal{M}_r(L_{\text{loc}}; \ell, f_{\text{true}}, \mathcal{F}^{\text{loc}})$, which is effective for small $r$. This does not decay with $|\mathcal{R}^X|$, hence becomes worse with increasing $r$ and complements the above standard.

### 3.3. Illustrative examples

Assume the $\mathcal{F}^{\text{global}}$ admits $q^x$-th order derivative in the region $\mathcal{B}(x, r)$. Then a natural choice for $|\mathcal{F}^x|$ is the set of multivariate polynomial functions of degree $q^x$, namely $\mathcal{P}(q^x)$, for some $q^x \geq 1$. The $L_1$ approximation error between $\mathcal{F}^x \equiv \mathcal{P}(q^x)$ and $\mathcal{F}^{\text{global}}$ can be quantified using the remainder in Taylor's approximation. This remainder typically grows as $C(\mathcal{F}^{\text{global}}, q^x) r^{(q^x+1)}$ for our choice of radius $r$ for the neighborhood, where $C(\mathcal{F}^{\text{global}}, q^x)$ depends on the function class and the degree. Therefore, we have $\varepsilon_{\mathcal{X}} \leq C(\mathcal{F}^{\text{global}}, q^x) r^{(q^x+1)}$.

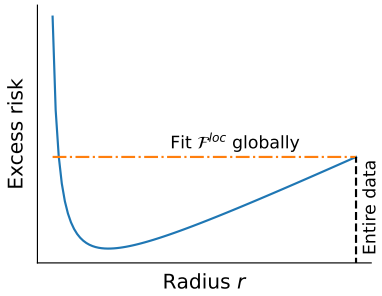



*Figure 2.* Behavior of excess risk of local ERM

**Local linear models.** Let us consider this setting where $\mathcal{F}^{\text{loc}}$ is the class of linear classifiers in $d$-dimension. The error in approximating $\mathcal{F}^x = \mathcal{P}(q^x)$ for any $q^x > 1$ with a linear classifier in the $\mathcal{B}(x, r)$ neighborhood for any $x \in \mathcal{X}$ is bounded by $\varepsilon_{\text{loc}} = \Theta(r^2)$. Therefore, the term (I) admits the bound $O(r^2)$. The generalization term varies as $O(1/\sqrt{N(r, \delta)})$. For $r \geq \Omega(n^{-1/2d} \log(n)^{1/2})$ and $\delta = n^{-1/2d}$ then $N(r, \delta) = \Omega(\sqrt{n^{(2d-1)/2d} r^d})$. Combining this we obtain:

$$\text{Excess Risk} \leq \underbrace{O\big(r^2\big)}_{\text{(I)}} + \underbrace{O\big(r^{\min\{\alpha_{\text{true}},1\}}\big)}_{\text{(II)}} + \underbrace{O\Big(\frac{d}{n^{(2d-1)/2d} r^{d/2}} + \frac{r^{\min\{\alpha_{\text{true}},1\}}}{n^{(2d-1)/4d} r^{d/2}} + \frac{1}{n^{1/2d}}\Big)}_{\text{(III)}}.$$

For $r = n^{-1/2d} \log(n)^{1/2}$ the excess risk bound is $O(n^{-1/2d} \log(n)^{1/2})$, where the bottleneck comes from the term (II), i.e., the sample vs retrieved risk. This is depicted in Fig. 2. Moreover, when the data has multivariate Gaussian distribution we have the term (II) scale as $O(r^2)$, leading to excess risk of $O(n^{-1/d} \log(n)^{1/2})$. However, global ERM with linear classifiers increases the approximation error considerably. In particular, now approximation error becomes a constant $O(\text{diam}(\mathcal{X})^2)$, and dwarfs the generalization that decreases as $O(1/\sqrt{n})$.

**Feed-forward classifiers.** As another concrete example we study the setting where $\mathcal{F}^{\text{loc}}$ is a the class of fully connected deep neural networks (FC-DNN). We have $f_y(\cdot)$ to be an $L$ layer feed-forward network with 1-Lipschitz nonlinearities (Bartlett et al., 2017). Let, for layers $l = 1$ to $L$, the dimension of the weight matrix be $(d_l \times d_{l-1})$ with $d_L = |\mathcal{Y}|$. Also, let $b_l$ and $s_l$ be the $\ell_{2,1}$ norm and spectral norm upper bounds for layer $l$ weight matrix, respectively, with $b_l/s_l \leq \kappa$. We define $d_{\max} = \max_{l \in [L]} d_l$ and let $\tilde{B} = \max_{x \in \mathcal{X}} \|x\|_2 \prod_{l=1}^{L} s_l$.

*Approximation Error $\varepsilon_{\text{loc}}$:* For bounding $\varepsilon_{\text{loc}}$ in (6), we require $L_1$ error of $\mathcal{F}^{\text{loc}}$ in approximating polynomials of degree $q_{\max} = \max_{x \in \mathcal{X}} q^x$. An FC-DNN that can approximate polynomials with degree at most $q_{\max}$ upto $L_1$ error $\varepsilon_{\text{loc}}$ has (see, Theorem 9 in Liang & Srikant (2016))[2]

$$\text{depth, } L = O\big(q_{\max} + \log(d q_{\max} C'(\mathcal{F}^{\text{global}}, q_{\max})/\varepsilon_{\text{loc}})\big),$$
$$\text{width, } d_{\max} = O\big(d \log(d q_{\max} C'(\mathcal{F}^{\text{global}}, q_{\max})/\varepsilon_{\text{loc}})\big).$$

Here, $C'(\mathcal{F}^{\text{global}}, q^x)$ is a constant independent of $r$ and $\epsilon$.

*Rademacher complexity:* We now bound the term $\mathbb{E}_{(X,Y)\sim \mathsf{D}}[\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X, Y)\big) | |\mathcal{R}^X| > N(r, \delta)]$ for this class. Following (Bartlett et al., 2017), for some universal constant $C'' > 0$ and any $\delta > 0$, we can bound the term as

$$C''\big(\tfrac{L_\ell \tilde{B} \sqrt{\kappa} \ln(d_{\max}) L^{3/4} \ln(L_\ell \tilde{B} \sqrt{n})^{3/2}}{\sqrt{N(r,\delta)}} + 2\delta \tilde{B}\big).$$

We now provide an excess risk bound when $\mathcal{F}^{\text{loc}}$ is the class of FC-DNN. Let $r \geq \Omega(n^{-1/2d} \log(n)^{1/2})$ and $\delta = n^{-1/2d}$. Then, $N(r, \delta) = \Omega(\sqrt{n^{(2d-1)/2d} r^d})$. Now, by setting $\varepsilon_{\text{loc}} = r^{(q_{\max}+1)}$, it follows from Thm. 3.7 that

$$\text{Excess Risk} \leq \underbrace{O\big(r^{(q_{\max}+1)}\big)}_{\text{(I)}} + \underbrace{O\big(r^{\min\{\alpha_{\text{true}},1\}}\big)}_{\text{(II)}} + \underbrace{O\Big(\tfrac{q_{\max}^{3/4} \ln(d q_{\max}/r)^{3/4} \ln(n)^{3/2}}{n^{(2d-1)/2d} r^{d/2}} + \tfrac{r^{\min\{\alpha_{\text{true}},1\}}}{n^{(2d-1)/4d} r^{d/2}} + \tfrac{1}{n^{1/2d}}\Big)}_{\text{(III)}}.$$

With $r = n^{-1/2d} \log(n)^{1/2}$, the excess risk is bounded as $O(n^{-1/2d} \log(n)^{1/2})$. Again (II) is the bottleneck. This bottleneck can be improved for multivariate Gaussian distribution with excess risk $O(n^{-1/d} \log(n)^{1/2}$.

Note that, it's also worth comparing local-ERM with conventional (non-local) ERM. Under the local structure condition (Sec. 2.2), one would utilize a simple $\mathcal{F}^{\text{loc}}$ for local-ERM. This would correspond to the Rademacher complexity term in Thm. 3.7 being small. In contrast, the generalization bound for the traditional (non-local) ERM approach would depend on the Rademacher complexity of a function class $\mathcal{F}^{\text{global}}$ that can achieve a low approximation error on the *entire domain*. Such a function class (even under the local structure assumption) would be much more complex than

[2] Although width is not explicitly mentioned in (Liang & Srikant, 2016), it can be inferred from the constructions.

$\mathcal{F}^{\text{loc}}$, resulting in a large Rademacher complexity. For the right design choice of $r$, and $\mathcal{F}^{\text{loc}}$, the approximation error increase of local-ERM can be offset by large generalization error of $\mathcal{F}^{\text{global}}$. As a consequence, local ERM with simple function class $\mathcal{F}^{\text{loc}}$ can outperform (non-local) ERM with a complex class $\mathcal{F}^{\text{global}}$.

### 3.4. Endowing local ERM with global representations

Note that the local ERM method takes a somewhat myopic view and does not aim to learn a global hypothesis that (partially or entirely) explains the entire data distribution. Such an approach may potentially result in poor performance in those regions of input domains that are not well represented in the training set. Here, we explore a two-stage approach leveraging the global pattern present in the training data to address this apparent shortcoming of local ERM.

Given training data $\mathcal{S}$ and a simple function class $\mathcal{G}^{\text{loc}} : \mathbb{R}^d \to \mathbb{R}^{|\mathcal{Y}|}$, the first stage involves learning a $d$-dimensional feature map $\Phi_{\mathcal{S}} : \mathcal{X} \to \mathbb{R}^d$ that simultaneously ensures good representation for the entire data distribution (Radford et al., 2021; Grill et al., 2020; Cer et al., 2018; Reimers & Gurevych, 2019). Subsequently, given a test instance $x$ and its retrieved neighboring points $\mathcal{R}^x = \{(x'_j, y'_j)\} \subseteq \mathcal{S}$, one employs local ERM with the function class:

$$\mathcal{F}_{\Phi_{\mathcal{S}}} = \{x \mapsto g \circ \Phi_{\mathcal{S}}(x) : g \in \mathcal{G}^{\text{loc}}\}. \tag{16}$$

At this point, it is tempting to invoke the proof strategy outlined following Lem. 3.5, with $\mathcal{F}^{\text{loc}}$ replaced with $\mathcal{F}_{\Phi_{\mathcal{S}}}$ to characterize the performance of the aforementioned two-stage method. Note that one can indeed bound the first two terms appearing in Lem. 3.5 for the two-stage method as well. However, bounding the third term that corresponds to generalization gap for local ERM becomes challenging as $\mathcal{F}_{\Phi_{\mathcal{S}}}$ depends on $\mathcal{S}$ via the global representation $\Phi_{\mathcal{S}}$ learned in the first stage. Interestingly, Foster et al. (2019) explored a general framework to address such dependence for standard (non retrieval-based) learning. In fact, as an instantiation of their general framework, Foster et al. (2019, Sec. 5.4) consider the ERM in feature space defined by a representation. We employ their techniques to obtain the following result on the generalization gap for local ERM with $\mathcal{F}_{\Phi_{\mathcal{S}}}$.

**Proposition 3.8.** *Assuming the representation learned during the first stage is $\Delta$-sensitive, i.e., for $\mathcal{S}$ and $\mathcal{S}'$ that differ in a single example, we have $\|\Phi_S(x) - \Phi_{S'}(x)\| \leq \Delta \; \forall x \in \mathcal{X}$. Furthermore, we assume that each $g \in \mathcal{G}^{\text{loc}}$ (cf. 16) is $L$-Lipschitz, the loss $\ell : \mathbb{R}^{|\mathcal{Y}|} \times |\mathcal{Y}| \to \mathbb{R}$ is $L_{\ell,1}$-Lipschitz w.r.t. $\|\cdot\|_\infty$-norm in the first argument, and $\ell$ is bounded by $M_\ell$. Then, following holds with probability at least $1-\delta$:*

$$\sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}}}} \left| \mathop{\mathbb{E}}_{(X',Y') \sim \mathsf{D}^{x,r}} [\ell(f(X'), Y')] - \hat{R}^x_\ell(f) \right| \leq \left(M_\ell + 2\Delta L L_{\ell,1} |\mathcal{R}^x|\right) \sqrt{\log(1/\delta)/2|\mathcal{R}^x|} + \mathop{\mathbb{E}}_{\mathcal{R}^x \sim \mathsf{D}^{x,r}} \left[ \sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}}}} \left| R_\ell(f) - \hat{R}^x_\ell(f) \right| \right]. \tag{17}$$

*Furthermore,*

$$\mathop{\mathbb{E}}_{\mathcal{R}^x \sim \mathsf{D}^{x,r}} \left[ \sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}}}} \left| R_\ell(f) - \hat{R}^x_\ell(f) \right| \right] \leq 2\mathfrak{R}^\diamond(\ell \circ \mathcal{F}_{\Phi_{\mathcal{S}}}), \tag{18}$$

*where $\ell \circ \mathcal{F}_{\Phi_{\mathcal{S}}} = \{(x,y) \mapsto \ell(f(x), y) : f \in \mathcal{F}_{\Phi_{\mathcal{S}}}\}$ and $\mathfrak{R}^\diamond$ denotes the Rademacher complexity of data dependent hypothesis sets (Foster et al., 2019).*

We defer the proof of Prop. 3.8 and necessary background on Foster et al. (2019) to Appendix D.

As a potential advantage of utilizing a global representation with local ERM, one can realize high-performance local learning with an even simpler function class. For example, it's a common approach to only train a linear classifier on learned representations. Furthermore, a high-quality global representation can ensure good performance for those local regions that are not well represented in the training set. We leave a formal treatment of these topics for a longer version of this manuscript.

## 4. Classification in extended feature space

Next, we focus on a family of retrieval-based methods that directly learn a scorer to map an input instance and its neighboring labeled instance to a score vector (cf. (7)). In fact, as discussed in Sec. 1, many successful modern instances of retrieval-based models such as REINA (Wang et al., 2022) and KATE (Liu et al., 2022) belong to this family. In this section, we provide the first rigorous treatment (to the best of our knowledge) for such models.

As introduced in Sec. 2.3, our objective is to learn a function $f : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^\star \to \mathbb{R}^{|\mathcal{Y}|}$. For a given instance $x$, such a function can leverage its neighboring set $\mathcal{R}^x \in (\mathcal{X} \times \mathcal{Y})^\star$ to improve the prediction on $x$. In this work, we restrict ourselves to a sub-family of such retrieval-based methods that first map $\mathcal{R}^x \sim \mathsf{D}^{x,r}$ to $\hat{\mathsf{D}}^{x,r}$ — an empirical estimate of the local distribution $\mathsf{D}^{x,r}$, which is subsequently utilized to make a prediction for $x$. In particular, the scorers of interest are of the form: $(x, \mathcal{R}^x) \mapsto f(x, \hat{\mathsf{D}}^{x,r})$, with

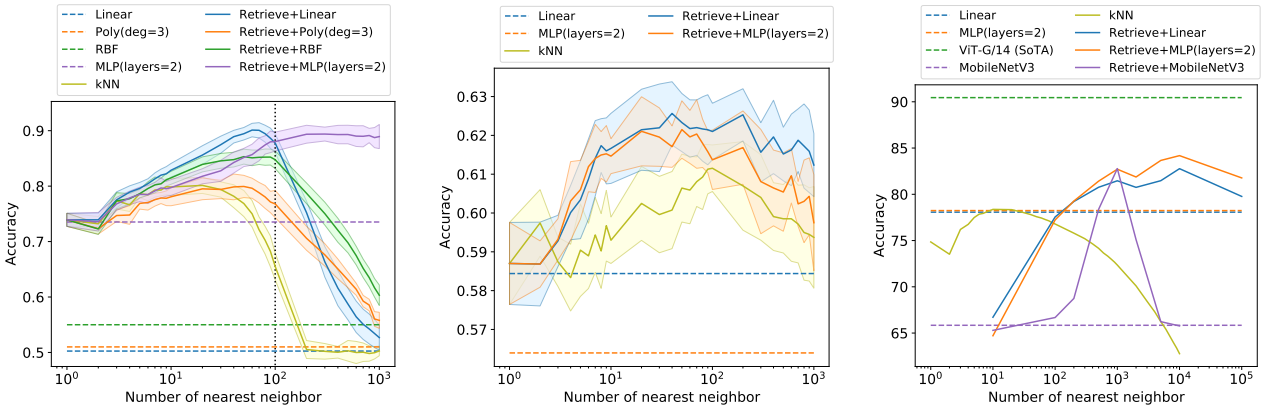
$$f(x, \hat{\mathsf{D}}^{x,r}) = \left(f_1(x, \hat{\mathsf{D}}^{x,r}), \ldots, f_{|\mathcal{Y}|}(x, \hat{\mathsf{D}}^{x,r})\right) \in \mathbb{R}^{|\mathcal{Y}|}.$$

Here, $f_y(x, \hat{\mathsf{D}}^{x,r})$ denotes the score assigned to the $y$-th class. Thus, assuming that $\Delta_{\mathcal{X}\times\mathcal{Y}}$ denotes the set of distribution over $\mathcal{X} \times \mathcal{Y}$, we restrict to a suitable function class in $\{f : \mathcal{X} \times \Delta_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|}\}$. Note that, given a surrogate loss $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ and scorer $f$, the empirical risk $\widehat{R}^{\text{ex}}_\ell(f)$ and population risk $R^{\text{ex}}_\ell(f)$ take the following form:

$$\widehat{R}^{\text{ex}}_\ell(f) = \frac{1}{n} \sum\nolimits_{i \in [n]} \ell\big(x_i, \hat{\mathsf{D}}^{x_i,r}\big)$$

and

$$R^{\text{ex}}_\ell(f) = \mathbb{E}_{(X,Y)\sim\mathsf{D}} \left[\ell\big(f(X, \mathsf{D}^{X,r}), Y\big)\right].$$




(a) Synthetic (b) CIFAR-10 (c) ImageNet

*Figure 3.* Performance of local ERM with size of retrieved set across models of different complexity.

Note that that the general framework for learning in the extended feature space $\widetilde{\mathcal{X}} := \mathcal{X} \times \Delta_{\mathcal{X}\times\mathcal{Y}}$ provides a very rich class of functions. In this paper, we focus on a specific form of learning methods in the extended feature space by using the kernel methods. The method as well as its analysis is obtained by adapting the work on utilizing kernel methods for domain generalization (Blanchard et al., 2011; Deshmukh et al., 2019).

In particular, we study generalization of a kernel-based classifier over $\widetilde{\mathcal{X}}$ learnt via regularized ERM. Due to space constraints, we present an informal version of our result below. See Appendix E for the precise statement (cf. Thm. E.4), necessary background, and detailed proof.

**Theorem 4.1** (Informal)**.** *Let $0 \le \delta \le 1$ and $N(r, \delta)$ be as defined in* (15)*. Then, under appropriate assumptions, with probability at least $1 - \delta$, we have*

$$\sup_{f\in\mathcal{F}} \left|\widehat{R}_\ell^{\mathrm{ex}}(f) - R_\ell^{\mathrm{ex}}(f)\right| \lesssim C_1 n^{-\frac{1}{2}} \left(1 + \log^{\frac{3}{2}} \sqrt{2} n|\mathcal{Y}|\right) + C_2\sqrt{\frac{\log(\frac{n}{\delta})}{N(r, \frac{\delta}{n})}} + C_3\sqrt{\frac{\log(\frac{1}{\delta})}{n}},$$

*where $\mathcal{F}$ is extended feature kernel function class; and $\widehat{R}_\ell^{\mathrm{ex}}(f)$ and $R_\ell^{\mathrm{ex}}(f)$ are empirical and population risks.*

Interestingly, the bound in Thm. 4.1 implies that the size of the retrieved set $\mathcal{R}^x$ (as captured by $N(r, \frac{\delta}{n})$) has to scale at least logarithmically in the size of the training set $n$ to ensure convergence.

## 5. Experiments

There have been numerous successful practical applications of retrieval-based models in the literature (e.g., Wang et al., 2022; Das et al., 2021). Here, we present a brief empirical study for such models in order to corroborate the benefits predicted by our theoretical results. We also present preliminary experiments to empirically verify the kernel based extended feature space-based approach in Appendix E.3.

**Task and dataset.** We perform experiments on both synthetic and real datasets, as summarized below. Further details are relegated to Appendix F.

*(i) Synthetic.* We consider a task of binary classification on a Gaussian mixture. Each mixture component is endowed with its local linear decision boundary. We randomly generate a train set of size $n = 10000$ in a 10-dimensional space. We use Euclidean distance for retrieval and perform a 10-fold cross-validation.

*(ii) CIFAR-10.* Next, we consider a task of binary classification on a *real data* for object detection. In particular, we consider a subset of CIFAR-10 dataset where we only restrict to images from "Cat" and "Dog" classes. We randomly partition the data into a train set of size $n = 10000$ points and remaining 2000 points for test. We use Euclidean distance for retrieval and do a 10-fold cross-validation.

*(iii) ImageNet.* Finally, we consider 1000-way classification task on ImageNet dataset. We use the standard train-test split with $n = 1281167$ training and 50000 test examples. Following standard practice in literature, we use unsupervised but globally learned features from ALIGN (Jia et al., 2021) to do image retrieval. This also showcases benefits of endowing local ERM with global representation (Sec. 3.4). Given large computational cost, we could only run each experiment once in this setting.

**Methods.** On all datasets, as baseline, we consider simple linear classifier and multi-layer perceptron (MLP) of two layers. For retrieval-based models, we consider each of the above methods as the local model to fit on retrieved data points via local ERM framework (Sec. 3). For synthetic datasets, we also considered support vector machines with polynomial kernel (of degree 3) and with radial basis function (RBF) kernel, both for baseline and local ERM. For ImageNet, we additionally consider the state-of-the-art (SoTA) single model published for this task, which is

from the most recent CVPR 2022 (Zhai et al., 2022), as a baseline. In addition, for ImageNet, we also consider the pretrain-finetune version of local ERM, where using the retrieved set we fine-tune a MobileNetV3 (Howard et al., 2019) model that has been pretrained on entire ImageNet.

**Observations.** In Fig. 3, we observe the tradeoff of varying the size of the retrieved set (as dictated by the neighborhood radius) on the performance of retrieval-based methods across all settings. We see that when the number of retrieved samples is small, local ERM has lower accuracy, this is due to large generalization error. When the size of the retrieved sample space is high, local ERM fails to minimize the loss effectively due to the lack of model capacity. We see that this effect being more pronounced for simpler function classes such as linear classifier as compared to MLP. In Fig. 3c, we see that, via local ERM with a small MobileNet-V3 model, we are able to achieve the top-1 accuracy of 82.78 whereas a regularly trained MobileNet-V3 model achieves the top-1 accuracy of only 65.80. Also the result is very competitive with SoTA of 90.45 with a *much larger model*. Thus, our empirical evaluation demonstrates the utility of retrieval-based models via simple local ERM framework. In particular, it allows small sized models to attain very high performance.

## 6. Related work and discussion

**Local polynomial regression.** Perhaps the most similar to our setup is the rich set of work on local polynomial regression, which has been around for a long time since the pioneering works of Stone (1977; 1980) . This line of work aims to fit a low-degree polynomial at each point in the data set based on a subset of data points. Such approaches gained a lot of attention as parametric regression was not adequate in various practical applications of the time. The performance of this approach critically depends on subset selected to locally fit the data. Towards this, various selection approaches have been considered: fixed bandwidth (Katkovnik & Kheisin, 1979), nearest neighbors (Cleveland, 1979), kernel weighted (Ruppert & Wand, 1994), and adaptive methods (Ruppert et al., 1995). All these work only analyze under mean squared error loss and do not handle classification nor provide finite sample generalization bounds, which we obtain in this work.

**Multi-task and meta learning.** At a surface level, our setup might resemble multi-task and meta learning frameworks. In multi-task learning, we are given the examples from $T$ tasks/distributions and the objective is to ensure good classification performance on all the tasks. In meta-learning, the setting is made harder by requiring good performance on a new target task. As a common approach in these settings, we learn a shared representation across the tasks and then learn a simple task-specific mapping on top of these learned shared features (Vilalta & Drissi, 2002, interalia). Theoretical investigations is quite limited: a few works study upper-bounds of generalization error in multi-task environments (Ben-David & Borbely, 2008; Ben-David et al., 2010; Pentina & Lampert, 2014; Amit & Meir, 2017), and even fewer in case of meta-learning (Balcan et al., 2019; Khodak et al., 2019; Du et al., 2020; Tripuraneni et al., 2021). However, most of these works assume linear or other simple class, whereas we consider general function class using kernel methods. It is not clear if the aforementioned representation based approach can apply to our setting because: each tasks have little overlap, very large number of tasks, and most importantly a priori an example belongs is not assigned to a task. Interestingly, in this work, we show that retrieval-based approach alleviate the needs to identify the task-membership. Here, we would like to highlight a contemporary work (Li et al., 2023) that studies in-context learning by Transformer models in a multi-task/meta-learning setting. In particular, this work relies on the notion of algorithm stability (Bousquet & Elisseeff, 2002) and presents generalization bounds for Transformers as in-context learners.

## 7. Conclusion and future direction

In this work, we initiate the development of a theoretical framework to study the statistical properties of retrieval-based modern machine learning models. Our treatment of an explicit local learning paradigm, namely local-ERM, establishes an approximation vs. generalization error tradeoff. This highlights the advantage realized by access to a retrieved set during classification as it enables good performance with much simpler (local) function classes. As for the retrieval-based models that leverage a retrieved set without explicitly performing local learning, we present a systematic study by considering a kernel-based classifier over extended feature space. Studying end-to-end retrieval-based models beyond kernel-based classification is a natural and fruitful direction for future work. It's also worth exploring if existing retrieval-based end-to-end models inherently perform *implicit* local learning via architectures such as Transformers.

## References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Amit, R. and Meir, R. Meta-learning by adjusting priors based on extended PAC-bayes theory. *arXiv preprint arXiv:1711.01244*, 2017.

Bai, Y. and Lee, J. D. Beyond linearization: On quadratic and higher-order approximation of wide neural networks. *arXiv preprint arXiv:1910.01619*, 2019.

Balcan, M.-F., Khodak, M., and Talwalkar, A. Provable guarantees for gradient-based meta-learning. In *Interna-*

*tional Conference on Machine Learning*, pp. 424–433. PMLR, 2019.

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

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Ben-David, S. and Borbely, R. S. A notion of task relatedness yielding provable multiple-task learning guarantees. *Machine learning*, 73(3):273–287, 2008.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1-2):151–175, 2010.

Blanchard, G., Lee, G., and Scott, C. Generalizing from several related classification tasks to a new unlabeled sample. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Blundell, C., Uria, B., Pritzel, A., Li, Y., Ruderman, A., Leibo, J. Z., Rae, J., Wierstra, D., and Hassabis, D. Model-free episodic control. *arXiv preprint arXiv:1606.04460*, 2016.

Bottou, L. and Vapnik, V. Local Learning Algorithms. *Neural Computation*, 4(6):888–900, 11 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.6.888.

Bousquet, O. and Elisseeff, A. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners, 2020.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Chen, M., Bai, Y., Lee, J. D., Zhao, T., Wang, H., Xiong, C., and Socher, R. Towards understanding hierarchical learning: Benefits of neural representations. *Advances in Neural Information Processing Systems*, 33:22134–22145, 2020.

Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Cramer, P. Alphafold2 and the future of structural biology. *Nature Structural & Molecular Biology*, 28(9):704–705, 2021.

Das, R., Zaheer, M., Thai, D., Godbole, A., Perez, E., Lee, J. Y., Tan, L., Polymenakos, L., and McCallum, A. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9594–9611, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.755.

Deshmukh, A. A., Lei, Y., Sharma, S., Dogan, U., Cutler, J. W., and Scott, C. A generalization error bound for multi-class domain generalization, 2019.

Döring, M., Györfi, L., and Walk, H. Rate of convergence of $k$-nearest-neighbor classification rule. *Journal of Machine Learning Research*, 18(227):1–16, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Du, S. S., Hu, W., Kakade, S. M., Lee, J. D., and Lei, Q. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

Fix, E. and Hodges, J. L. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, 57(3):238–247, 1989.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M., and Sridharan, K. Hypothesis set stability and generalization. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=flNZJ2eOet`.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances*

*in Neural Information Processing Systems*, 33:21271–21284, 2020.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., and Adam, H. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Iscen, A., Fathi, A., Schmid, C., Caron, M., and Bird, T. A memory transformer network for incremental learning. *arXiv preprint*, 2022.

Izacard, G., Lewis, P., Lomeli, M., Hosseini, L., Petroni, F., Schick, T., Dwivedi-Yu, J., Joulin, A., Riedel, S., and Grave, E. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Katkovnik, V. Y. and Kheisin, V. Dynamic stochastic approximation of polynomials drifts. *Avtomatika i Telemekhanika*, pp. 89–98, 1979.

Khodak, M., Balcan, M.-F. F., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Lei, Y., Dogan, Ü., Zhou, D.-X., and Kloft, M. Data-dependent generalization bounds for multi-class classification. *IEEE Transactions on Information Theory*, 65(5): 2995–3021, 2019.

Li, Y., Swersky, K., and Zemel, R. Generative moment matching networks. In *International conference on machine learning*, pp. 1718–1727. PMLR, 2015.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and stability in in-context learning. *arXiv preprint arXiv:2301.07067*, 2023.

Liang, S. and Srikant, R. Why deep neural networks for function approximation? *arXiv preprint arXiv:1610.04161*, 2016.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Liu, S., Liang, X., Liu, L., Shen, X., Yang, J., Xu, C., Lin, L., Cao, X., and Yan, S. Matching-cnn meets knn: Quasi-parametric human parsing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1419–1427, 2015.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Long, A., Yin, W., Ajanthan, T., Nguyen, V., Purkait, P., Garg, R., Blair, A., Shen, C., and van den Hengel, A. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6959–6969, 2022.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., Hesse, C., Jain, S., Kosaraju, V., Saunders, W., et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Pentina, A. and Lampert, C. A PAC-bayesian bound for lifelong learning. In *International Conference on Machine Learning*, pp. 991–999, 2014.

Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., and Blundell, C. Neural episodic control. In *International Conference on Machine Learning*, pp. 2827–2836. PMLR, 2017.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Ritter, S., Faulkner, R., Sartran, L., Santoro, A., Botvinick, M., and Raposo, D. Rapid task-solving in novel environments. *arXiv preprint arXiv:2006.03662*, 2020.

Ruppert, D. and Wand, M. P. Multivariate locally weighted least squares regression. *The annals of statistics*, pp. 1346–1370, 1994.

Ruppert, D., Sheather, S. J., and Wand, M. P. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432): 1257–1270, 1995.

Samarin, M., Roth, V., and Belius, D. On the empirical neural tangent kernel of standard finite-width convolutional neural network architectures. *arXiv preprint arXiv:2006.13645*, 2020.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In Hutter, M., Servedio, R. A., and Takimoto, E. (eds.), *Algorithmic Learning Theory*, pp. 13–31, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75225-7.

Steinwart, I. Adaptive density level set clustering. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 703–738. JMLR Workshop and Conference Proceedings, 2011.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Stone, C. J. Consistent nonparametric regression. *The annals of statistics*, pp. 595–620, 1977.

Stone, C. J. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pp. 1348–1360, 1980.

Tripuraneni, N., Jin, C., and Jordan, M. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

Vilalta, R. and Drissi, Y. A perspective view and survey of meta-learning. *Artificial intelligence review*, 18(2):77–95, 2002.

von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*, 2022.

Wang, S., Xu, Y., Fang, Y., Liu, Y., Sun, S., Xu, R., Zhu, C., and Zeng, M. Training data is more valuable than you think: A simple and effective method by retrieving from training data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3170–3179, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.226. URL https://aclanthology.org/2022.acl-long.226.

Yarotsky, D. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Zakai, A. and Ritov, Y. How local should a learning method be?. In *COLT*, pp. 205–216. Citeseer, 2008.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12104–12113, 2022.

Zhang, T. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5(Oct):1225–1251, 2004.

# A. Preliminaries

**Definition A.1** (Rademacher complexity)**.** *Given a sample $\mathcal{S} = \{z_i = (x_i, y_i)\}_{i\in[n]} \subset \mathcal{Z}$ and a real-valued function class $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$, the* empirical *Rademacher complexity of $\mathcal{F}$ with respect to $\mathcal{S}$ is defined as*

$$\mathfrak{R}_{\mathcal{S}}(\mathcal{F}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(z_i)\right], \tag{19}$$

*where $\boldsymbol{\sigma} = \{\sigma_i\}_{i\in[n]}$ is a collection of $n$ i.i.d. Bernoulli random variables. For $n \in \mathbb{N}$, the Rademacher complexity $\bar{\mathfrak{R}}_n(\mathcal{F})$ and* worst case *Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ are defined as follows.*

$$\bar{\mathfrak{R}}_n(\mathcal{F}) = \mathbb{E}_{\mathcal{S}\sim\mathsf{D}^n}\left[\mathfrak{R}_{\mathcal{S}}(\mathcal{F})\right], \quad \textit{and} \quad \mathfrak{R}_n(\mathcal{F}) = \sup_{\mathcal{S}\sim\mathcal{Z}^n}\mathfrak{R}_{\mathcal{S}}(\mathcal{F}). \tag{20}$$

**Definition A.2** (Covering Number)**.** *Let $\epsilon > 0$ and $\|\cdot\|$ be a norm defined over $\mathbb{R}^n$. Given a function class $\mathcal{F} : \mathcal{Z} \to \mathbb{R}$ and a collection of points $\mathcal{S} = \{z_i\}_{i\in[n]} \subset \mathcal{Z}$, we call a set of points $\{u_j\}_{j\in[m]} \subset \mathbb{R}^n$ an $(\epsilon, \|\cdot\|)$-cover of $\mathcal{F}$ with respect to $\mathcal{S}$, if we have*

$$\sup_{f\in\mathcal{F}}\min_{j\in[m]}\|f(\mathcal{S}) - u_j\| \le \epsilon, \tag{21}$$

*where $f(\mathcal{S}) = \big(f(z_1), \ldots, f(z_n)\big) \in \mathbb{R}^n$. The $\|\cdot\|$-covering number $\mathcal{N}_{\|\cdot\|}(\epsilon, \mathcal{F}, \mathcal{S})$ denotes the cardinally of the minimal $(\epsilon, \|\cdot\|)$-cover of $\mathcal{F}$ with respect to $\mathcal{S}$. In particular, if $\|\cdot\|$ is an normalized-$\ell_p$ norm ($\|v\| = (\frac{1}{dim(v)}\sum_{i=1}^{dim(v)}|v_i|^p)^{1/p}$), then we simply use $\mathcal{N}_p(\epsilon, \mathcal{F}, \mathcal{S})$ to denote the corresponding $\ell_p$-covering number.*

# B. Proofs for Section 3.2

## B.1. Proof of Lemma 3.5

Note that

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\ell(\hat{f}^X(X), Y) - \ell(f^*(X), Y)\Big] \\
&\quad \text{// We add and subtract loss of the local optimizer } f^{X,*}(\cdot) \text{ expected over } \mathsf{D}^{X,r} \\
&\quad = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\ell(\hat{f}^X(X), Y) - \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big] \\
&\qquad\qquad + \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big] - \ell(f^*(X), Y)\Big] \\
&\quad \text{// We add and subtract loss of the global optimizer } f^*(\cdot) \text{ expected over } \mathsf{D}^{X,r} \\
&\quad = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\ell(\hat{f}^X(X), Y) - \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big] \\
&\qquad\qquad + \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^*(X'), Y'\big)\big] - \ell(f^*(X), Y) \\
&\qquad\qquad + \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big] - \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\quad \text{// We group (1) local vs global optimizer, (2) global optimizer at } X \text{ vs expected over } \mathsf{D}^{X,r}, \\
&\quad \text{// and (3) ERM loss at } X \text{ vs local optimizer loss expected over } \mathsf{D}^{X,r} \\
&\quad = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\qquad + \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^*(X'), Y'\big)\big] - \ell(f^*(X), Y)\Big] \\
&\qquad + \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\ell(\hat{f}^X(X), Y) - \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big]\Big] \\
&\quad \text{// We add and subtract loss of the empirical optimizer } \hat{f}^X(\cdot) \text{ expected over } \mathsf{D}^{X,r} \\
&\quad = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\qquad + \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f^*(X'), Y'\big)\big] - \ell(f^*(X), Y)\Big]
\end{aligned}$$

$$+ \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\ell(\hat{f}^X(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(\hat{f}^X(X'), Y'\big)] + \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(\hat{f}^X(X'), Y'\big)] - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big]\Big]$$

// We (1) bound difference of loss at $X$ and loss expected over $\mathsf{D}^{X,r}$ by maximizing over function class,
// and (2) subtract empirical loss of empirical optimizer and add (larger) empirical loss of local optimizer

$$\begin{aligned}
&\leq \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{global}}} \big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f(X'), Y'\big)\big] - \ell(f(X), Y)\big|\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}} \big|\ell(f(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(f(X'), Y'\big)]|\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(\hat{f}^X(X'), Y'\big)] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(\hat{f}^X(x'), y'\big)\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(f^{X,*}(x'), y'\big) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big]\Big] \qquad (22)
\end{aligned}$$

// We (1) bound difference of empirical vs expected loss of empirical optimizer by maximizing over function class,

$$\begin{aligned}
&\leq \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{global}}} \big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f(X'), Y'\big)\big] - \ell(f(X), Y)\big|\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}} \big|\ell(f(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(f(X'), Y'\big)]|\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}} \Big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}[\ell\big(f(X'), Y'\big)] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(f(x'), y'\big)\Big|\Big] \\
&\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\Big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big)\big] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(f^{X,*}(x'), y'\big)\Big|\Big] \qquad (23)
\end{aligned}$$

□

### B.2. Proof of Theorem 3.7

As discussed in Sec. 3, the proof of Theorem 3.7 requires bounding three terms in Lemma 3.5. We now proceed to establishing the desired bounds.

**Local vs global loss.** The local vs global loss can bounded easily using the local regularity condition, and due to the fact that $\mathcal{F}^{\text{loc}} \approx \cup_x \mathcal{F}^x$. Let

$$f^{X,\text{loc}} = \arg\min_{f\in\mathcal{F}^X} \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f(X'), Y'\big)\big].$$

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\qquad \leq \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'), Y'\big) - \ell\big(f^{X,\text{loc}}(X'), Y'\big)\big]\Big] \\
&\qquad\quad + \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,\text{loc}}(X'), Y'\big) - \ell\big(f^*(X'), Y'\big)\big]\Big] \\
&\qquad \leq \varepsilon_{\text{loc}} + \varepsilon_{\mathcal{X}}.
\end{aligned}$$

**Global and local: Sample vs retrieved set risk.** The following lemma bounds the second term in Lemma 3.5. Recall the definition, for any $L > 0$,

$$\mathcal{M}_r(L; \ell, f_{\text{true}}, \mathcal{F}) = 2L_\ell \Big( Lr + \big(2\|\mathcal{F}\|_\infty - Lr\big) c_{\text{true}} \big(2L_{\text{true}} r\big)^{\alpha_{\text{true}}} \Big). \tag{24}$$

**Lemma B.1.** *Under Assumption 3.1, for a L-coordinate Lipschitz function class $\mathcal{F}$ with $\|\mathcal{F}\|_\infty := \sup_{x \in \mathcal{X}} \sup_{f \in \mathcal{F}} \|f(x)\|_\infty$ we have*

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim \mathsf{D}} \Big[ \sup_{f \in \mathcal{F}} \big| \ell(f(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}} [\ell\big(f(X'), Y'\big)] \big| \Big] \\
&\qquad \le 2L_\ell \Big( Lr + \big(2\|\mathcal{F}\|_\infty - Lr\big) c_{\text{true}} (2L_{\text{true}} r)^{\alpha_{\text{true}}} \Big).
\end{aligned}$$

*Proof.* We are given the example $(X, Y)$. Let us fix an arbitrary $f \in \mathcal{F}$, and any arbitrary example $(x', y')$ in the $r$ neighborhood of $X$.

We first bound the perturbation in $\gamma_f(\cdot)$ for a given label $\tilde{Y}$.

$$\begin{aligned}
|\gamma_f(X_1, \tilde{Y})) - \gamma_f(X_2, \tilde{Y})| &\le |f_{\tilde{Y}}(X_1) - \max_{s \neq \tilde{Y}} f_s(X_1) - f_{\tilde{Y}}(X_2) + \max_{s' \neq \tilde{Y}} f_{s'}(X_2)| \\
&\le |f_{\tilde{Y}}(X_1) - f_{\tilde{Y}}(X_2)| + |\max_{s \neq \tilde{Y}} f_s(X_1) - \max_{s' \neq \tilde{Y}} f_{s'}(X_2)| \\
&\le |f_{\tilde{Y}}(X_1) - f_{\tilde{Y}}(X_2)| + \max_{s \neq \tilde{Y}} |f_s(X_1) - f_s(X_2)| \\
&\le 2L \|X_1 - X_2\|_2
\end{aligned}$$

We can now proceed with bounding the loss.

$$\begin{aligned}
|\ell(f(X), Y) - \ell(f(x'), y')| &= |\ell(\gamma_f(X, Y)) - \ell(\gamma_f(x', y'))| \\
&\le L_\ell |\gamma_f(X, Y) - \gamma_f(x', y')| \\
&\le \begin{cases} 4L_\ell \|f\|_\infty; Y \neq y' \\ 2L_\ell Lr; Y = y' \end{cases}
\end{aligned}$$

Under Assumption 3.1, if we have $\gamma_{f^{\text{true}}}(X, Y) > 2L_{\text{true}} r$, then following the above argument we have $\gamma_{f^{\text{true}}}(X', Y) > 0$, thus $Y$ is the true label of $X'$. In other words, $\gamma_{f^{\text{true}}}(X, Y) > 2L_{\text{true}} r$ imply for any $X'$ in the $r$ neighborhood of $X$ its true label $Y' = Y$.

$$\begin{aligned}
&|\ell(f(X), Y) - \ell(f(x'), y')| \\
&\qquad \le 2L_\ell Lr \mathbb{1}(\gamma_{f^{\text{true}}}(X, Y) > 2L_{\text{true}} r) + 4L_\ell \|f\|_\infty \mathbb{1}(\gamma_{f^{\text{true}}}(X, Y) \le 2L_{\text{true}} r) \\
&\qquad \le 2L_\ell Lr + 2L_\ell \big(2\|f\|_\infty - Lr\big) \mathbb{1}(\gamma_{f^{\text{true}}}(X, Y) \le 2L_{\text{true}} r)
\end{aligned}$$

As $(x', y')$ was an arbitrary $r$-neighbor, we have

$$\begin{aligned}
&|\ell(f(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}} \ell(f(X'), Y')| \\
&\qquad \le \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}} |\ell(f(X), Y) - \ell(f(X'), Y')| \\
&\qquad \le 2L_\ell Lr + 2L_\ell \big(2\|f\|_\infty - Lr\big) \mathbb{1}(\gamma_{f^{\text{true}}}(X, Y) \le 2L_{\text{true}} r)
\end{aligned}$$

Furthermore, as $f$ was arbitrary, we have

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} |\ell(f(X), Y) - \mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}} \ell(f(X'), Y')| \\
&\qquad \le \sup_{f \in \mathcal{F}} 2L_\ell Lr + 2L_\ell \big(2\|f\|_\infty - Lr\big) \mathbb{1}(\gamma_{f^{\text{true}}}(X, Y) \le 2L_{\text{true}} r)
\end{aligned}$$

$$= 2L_\ell Lr + 2L_\ell\big(2\|\mathcal{F}\|_\infty - Lr\big)\mathbb{1}(\gamma_{f^{\text{true}}}(X,Y) \le 2L_{\text{true}}r).$$

Note $f^{\text{true}}$ is independent of $f$, which was used in the derivation of above inequalities. Taking expectation over $(X,Y)$, and using the margin condition as given in assumption 3.1 we obtain

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\sup_{f\in\mathcal{F}}|\ell(f(X),Y) - \mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\ell(f(X'),Y')|\Big] \\
&\qquad = 2L_\ell Lr + 2L_\ell\big(2\|\mathcal{F}\|_\infty - Lr\big)\mathbb{P}_{(X,Y)\sim\mathsf{D}}\Big[\gamma_{f^{\text{true}}}(X,Y) \le 2L_{\text{true}}r\Big] \\
&\qquad \le 2L_\ell Lr + 2L_\ell\big(2\|\mathcal{F}\|_\infty - Lr\big)c_{\text{true}}(2L_{\text{true}}r)^{\alpha_{\text{true}}} = \mathcal{M}_r(L;\ell,f_{\text{true}},\mathcal{F}).
\end{aligned}$$

□

Plugging in the Lipschitz bounds for the function classes $\mathcal{F}^{\text{loc}}$ and $\mathcal{F}^{\text{global}}$ in the above lemma bounds the second term.

An alternative way of bounding the risk difference is as follows:

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}\big[\ell\big(f(X'),Y'\big)\big] - \ell(f(X),Y)\Big] \\
&= \int_{x\in\mathcal{X}}\big(\int_{x'\in B(x,r)\cap\mathcal{X}}\tfrac{h(x')\rho_{\mathsf{D}}(x')}{\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}]}dx'\big)\rho_{\mathsf{D}}(x)dx - \int_{x\in\mathcal{X}}h(x)\rho_{\mathsf{D}}(x)dx \\
&= \int_{x'\in\mathcal{X}}h(x')\big(\int_{x\in B(x',r)\cap\mathcal{X}}\tfrac{\rho_{\mathsf{D}}(x)}{\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}]}dx\big)\rho_{\mathsf{D}}(x')dx' - \int_{x\in\mathcal{X}}h(x)\rho_{\mathsf{D}}(x)dx \\
&= \int_{x'\in\mathcal{X}}h(x')\big(\int_{x\in B(x',r)\cap\mathcal{X}}\tfrac{\rho_{\mathsf{D}}(x)}{\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}]}dx - 1\big)\rho_{\mathsf{D}}(x')dx' \\
&\le h_{\max}\int_{x'\in\mathcal{X}}\Big|\int_{x\in B(x',r)\cap\mathcal{X}}\tfrac{\rho_{\mathsf{D}}(x)}{\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}]}dx - 1\Big|\rho_{\mathsf{D}}(x')dx' \\
&\le h_{\max}\int_{x'\in\mathcal{X}}\max\big(\big|\tfrac{1}{1\pm c_{\text{wdc}+}r^{\alpha_{\text{wdc}+}}} - 1\big|\big)\rho_{\mathsf{D}}(x')dx' = \frac{h_{\max}c_{\text{wdc}+}r^{\alpha_{\text{wdc}+}}}{1 - c_{\text{wdc}+}r^{\alpha_{\text{wdc}+}}}.
\end{aligned}$$

We can express $\ell(f(X),Y) = h(X)$ because $Y$ is a deterministic function of $X$.

Under Assumption 3.4, with constants $c_{wdc+}$ and $\alpha_{wdc+}$, recall that

$$\Big|\frac{\mathbb{P}_{X'\sim\mathsf{D}}[\mathrm{d}(X',x)\le r]}{\rho_{\mathsf{D}}(x)vol_d(r)} - 1\Big| \le c_{\text{wdc}+}r^{\alpha_{\text{wdc}+}}.$$

Plugging this in gives us the final inequality.

**Example:** Let us consider the term $\frac{\mathbb{P}_{\mathsf{D}}[B(x',r)]}{\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}]}$ for $\mathsf{D}$ being multivariate Gaussian $N(\mu,\Sigma)$. Let $vol_d(r)$ imply the volume, and $S_d(r)$ the surface area of a $d$-sphere of radius $r$ in dimension $d$.

$$\begin{aligned}
\mathbb{P}_{\mathsf{D}}[B(x,r)\cap\mathcal{X}] &= \int_{z\in B(x,r)\cap\mathcal{X}}\tfrac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\tfrac{1}{2}(z-\mu)^T\Sigma^{-1}(z-\mu))dz \\
&= \tfrac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))\int_{z\in B(x,r)\cap\mathcal{X}}\exp(-\tfrac{1}{2}(z+x-2\mu)^T\Sigma^{-1}(z-x))dz \\
&= \rho_{\mathsf{D}}(x)\int_{u\in B(0,r)}\exp(-\tfrac{1}{2}(u+2(x-\mu))^T\Sigma^{-1}u)du \\
&= \rho_{\mathsf{D}}(x)\big(vol_d(r) - c\int_{u\in B(0,r)}(u+2(x-\mu))^T\Sigma^{-1}udu\big) \text{ for some } c\in[1/4,1/2] \\
&= \rho_{\mathsf{D}}(x)\big(vol_d(r) - c\int_{u\in B(0,r)}(c_1\|u\|_2^2 + 2(x-\mu)^T\Sigma^{-1}u)du\big) \text{ for some } c_1\in[\lambda_{\min}(\Sigma^{-1}),\lambda_{\max}(\Sigma^{-1})]
\end{aligned}$$

We have used $exp(-x)\in(1-x,1-x/2)$ for $x\le 1.59$. Also, $\frac{u^T\Sigma^{-1}u}{\|u\|_2^2}\in[\lambda_{\min}(\Sigma^{-1}),\lambda_{\max}(\Sigma^{-1})]$.

Then using polar coordinate transform we obtain

$$\int_{u\in B(0,r)} \|u\|_2^2 du = \int_0^r l^2 S_d(l) dl = \int_0^r l^2 \tfrac{d\, l^{d-1} \pi^{d/2}}{\Gamma(1+d/2)} dl = r^{d+2} \tfrac{d\, \pi^{d/2}}{(d+2)\Gamma(1+d/2)} = vol_d(r) \tfrac{dr^2}{d+2}.$$

Let $\xi = 2(x-\mu)\Sigma^{-1}$. We want to integrate $\xi^T u$ over $B(0,r)$. Through a somewhat different polar transform where the polar axis is parallel to $\xi$ and the angle of $u$ and $\xi$ is $\theta$ we can do the integral as follows for $d \geq 2$.

$$\int_{u\in B(0,r)} \xi^T u du = \int_{l=0}^r \int_{\theta=0}^{\pi} |\xi| l \cos(\theta) S_{d-1}(l) \sin^{d-2}(\theta) d\theta dl = \tfrac{d\pi^{d/2}}{\Gamma(1+d/2)} \int_{l=0}^r |\xi| l^{d-1} dl \underbrace{\int_{\theta=0}^{\pi} \cos(\theta) \sin^{d-2}(\theta) d\theta}_{=0} = 0.$$

Substituting, these values in the above inequality we get

$$\mathbb{P}_{\mathsf{D}}[B(x,r) \cap \mathcal{X}] = \rho_{\mathsf{D}}(x) vol_d(r)\big(1 - c_2 r^2\big), \text{ for some } c_2 = [\tfrac{d\lambda_{\min}(\Sigma^{-1})}{4(d+2)}, \tfrac{d\lambda_{\max}(\Sigma^{-1})}{2(d+2)}].$$

Therefore, the difference of Retrieved vs Sample risk for multi-variate Gaussian is bounded as

$$\frac{l_{\max}(f) dr^2}{(d+2)\lambda_{\max}(\Sigma)} \text{ for } r \leq \sqrt{(d+2)\lambda_{\max}(\Sigma)}.$$

**Generalization of local ERM.** Recall the function class $\mathcal{G}(X,Y) = \{\ell(\gamma_f(\cdot,\cdot)) - \ell(\gamma_f(X,Y)) : f \in \mathcal{F}^{\text{loc}}\}$. Here $\mathcal{G}(X,Y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. Note that the function class is parameterized by $(X,Y)$. Let us define some quantities of the function class on a set $S \subseteq \mathcal{X} \times \mathcal{Y}$ as

$$\mathcal{G}_{\max}((X,Y);S) = \sup_{g\in\mathcal{G}(X,Y)} \sup_{(x',y')\in S} |g(x',y')|$$

By centering each function $f \in \mathcal{F}^{\text{loc}}$ at the point $(X,Y)$ we can transform the generalization over the function class $\mathcal{F}^{\text{loc}}$, to the generalization over the function class $\mathcal{G}(X,Y)$. In particular, we have

$$\begin{aligned}
&\mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}} \Big|\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}[\ell\big(f(X'),Y'\big)] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(f(x'),y'\big)\Big|\Big] \\
&\leq \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\sup_{f\in\mathcal{F}^{\text{loc}}} \Big|\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}[\ell\big(f(X'),Y'\big) - \ell\big(f(X),Y\big)] \\
&\qquad - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} \ell\big(f(x'),y'\big) - \ell\big(f(X),Y\big)\Big|\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + 4\delta L_\ell \|\mathcal{F}^{\text{loc}}\|_\infty \\
&= \mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\sup_{g\in\mathcal{G}(X,Y)} \Big|\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}[g(X',Y')] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} g(x',y')\Big|\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + 4\delta L_\ell \|\mathcal{F}^{\text{loc}}\|_\infty.
\end{aligned}$$

We next state a standard result of learning theory that bounds the final term using the Rademacher complexity of the function class $\mathcal{G}(X,Y)$ (Shalev-Shwartz & Ben-David, 2014).

**Lemma B.2** (Adapted from Theorem 26.5 in Shalev-Shwartz & Ben-David (2014).)**.** *For any $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ and a neighborhood set $\mathcal{R}^X$, and any function $g \in \mathcal{G}(X,Y)$, for each $\delta > 0$ with probability at least $(1-\delta)$ the following holds*

$$\Big|\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}[g\big(X',Y'\big)] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} g\big(x',y'\big)\Big| \leq 2\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big) + 4\mathcal{G}_{\max}((X,Y);\mathcal{R}^X)\sqrt{\frac{2\ln(4/\delta)}{|\mathcal{R}^X|}}.$$

Taking expectation with respect to $(X,Y)$, we obtain

$$\mathbb{E}_{(X,Y)\sim\mathsf{D}}\Big[\sup_{g\in\mathcal{G}(X,Y)} \Big|\mathbb{E}_{(X',Y')\sim\mathsf{D}^{X,r}}[g\big(X',Y'\big)] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X} g\big(x',y'\big)\Big|\Big||\mathcal{R}^X| \geq N(r,\delta)\Big]$$

$$
\begin{aligned}
&\leq 2\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + \\
&\quad 4\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathcal{G}_{\max}((X,Y);\mathcal{R}^X)\sqrt{\frac{2\ln(4/\delta)}{|\mathcal{R}^X|}}\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + 4\delta L_\ell \|\mathcal{F}^{\mathrm{loc}}\|_\infty \\
&\leq 2\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + \\
&\quad 4\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathcal{G}_{\max}((X,Y);\mathcal{R}^X)\Big||\mathcal{R}^X| \geq N(r,\delta)\Big]\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sqrt{\frac{2\ln(4/\delta)}{|\mathcal{R}^X|}}\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + 4\delta L_\ell \|\mathcal{F}^{\mathrm{loc}}\|_\infty \\
&\leq 2\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)\Big||\mathcal{R}^X| \geq N(r,\delta)\Big] + 4\mathcal{M}_r(L_{\mathrm{loc}};\ell,f_{\mathrm{true}},\mathcal{F}^{\mathrm{loc}})\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sqrt{\frac{2\ln(4/\delta)}{N(r,\delta)}}\Big] + 4\delta L_\ell \|\mathcal{F}^{\mathrm{loc}}\|_\infty.
\end{aligned}
$$

In the first inequality, with probability $(1-\delta)$ we apply the bound from Lemma B.2, whereas we use the bound $4L_\ell\|\mathcal{F}^{\mathrm{loc}}\|_\infty$ with remaining probability $\delta$. Also from the proof of Lemma B.1 we have that

$$
\mathcal{G}_{\max}((X,Y);\mathcal{R}^X) \leq 2L_\ell\Big(Lr + \big(\max\{Lr, 2\|\mathcal{F}^{\mathrm{loc}}\|_\infty\} - Lr\big)\mathbb{1}\big(\gamma_{f^{\mathrm{true}}}(X,Y) \leq 2L_{\mathrm{true}}r\big)\Big).
$$

Taking expectation with respect to $\mathsf{D}$ completes the bound. While taking expectation we crucially use the fact that $\gamma_{f^{\mathrm{true}}}(X,Y)$ is independent of $|\mathcal{R}^X|$ to arrive at the $\mathcal{M}_r(L_{\mathrm{loc}};\ell,f_{\mathrm{true}},\mathcal{F}^{\mathrm{loc}})$ bound.

**Central absolute moment of $f^{X,*}$.** As the function $f^{X,*}$ is *fixed* using centering, and then Hoeffding bound, we can directly bound the remaining term. We have with probability at least $(1-\delta)$

$$
\begin{aligned}
&\Big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'),Y'\big)\big] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}\ell\big(f^{X,*}(x'),y'\big)\Big| \\
&= \Big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'),Y'\big) - \ell\big(f^{X,*}(X),Y\big)\big] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}\ell\big(f^{X,*}(x'),y'\big) - \ell\big(f^{X,*}(X),Y\big)\Big| \\
&\leq \mathcal{G}_{\max}((X,Y);\mathcal{R}^X)\sqrt{\frac{\ln(2/\delta)}{|\mathcal{R}^X|}}
\end{aligned}
$$

Taking expectation similar to the previous case we obtain,

$$
\begin{aligned}
&\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\Big|\mathbb{E}_{(X',Y')\sim \mathsf{D}^{X,r}}\big[\ell\big(f^{X,*}(X'),Y'\big)\big] - \frac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}\ell\big(f^{X,*}(x'),y'\big)\Big|\Big] \\
&\leq \mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\min\{4L_\ell\|\mathcal{F}^{\mathrm{loc}}\|_\infty, \mathcal{G}_{\max}((X,Y);\mathcal{R}^X)\sqrt{\frac{\ln(2/\delta)}{|\mathcal{R}^X|}}\}\Big] \\
&\leq \mathcal{M}_r(L_{\mathrm{loc}};\ell,f_{\mathrm{true}},\mathcal{F}^{\mathrm{loc}})\mathbb{E}_{(X,Y)\sim \mathsf{D}}\Big[\sqrt{\frac{\ln(2/\delta)}{N(r,\delta)}}\Big] + 4\delta L_\ell\|\mathcal{F}^{\mathrm{loc}}\|_\infty.
\end{aligned}
$$

Here, we use the fact that $\gamma_{f^{\mathrm{true}}}(X,Y)$ is independent of $|\mathcal{R}^X|$. This concludes the proof of Theorem 3.7.

### B.3. Bounding the Rademacher complexity $\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)$

We now derive bounds on the Rademacher complexity of the class $\mathcal{G}(X,Y)$. We use the covering number based bounds for that purpose. We then start by relating it to the covering number of the $\mathcal{F}^{\mathrm{loc}}$ function class. Finally, we provide a bound on the class of functions residing in bounded norm Reproducing Kernel Hilbert Space.

We will use $\mathcal{G}_{\max}(X,Y)$ instead of $\mathcal{G}_{\max}((X,Y);\mathcal{R}^X)$ when the context is clear. Similar to $\mathcal{G}(X,Y)$, we define the function class $\mathcal{G} = \{\ell(\gamma_f(\cdot,\cdot)) : f \in \mathcal{F}^{\mathrm{loc}}\}$ which does not depend on the locality centered around $(X,Y)$. On a set $S \subseteq \mathcal{X}\times\mathcal{Y}$ we can define $\mathcal{G}_{\max}(S) = \sup_{g\in\mathcal{G}}\sup_{(x',y')\in S}|g(x',y')|$.

**Lemma B.3.** *Under Assumption 3.1 we have for any retrieved set within radius $r$ of $X$, $\mathcal{R}^X$, for any $p \geq 1$*

$$
\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big)
$$

$$\leq \begin{cases} \mathcal{G}_{\max}(X,Y) \\ \inf_{\epsilon\in[0,\mathcal{G}_{p,\max}(X,Y)/2]} \left(4\epsilon + \frac{12}{\sqrt{|\mathcal{R}^X|}} \int_\epsilon^{\mathcal{G}_{p,\max}(X,Y)/2} \sqrt{\log\left(\frac{2\mathcal{G}_{\max}}{\nu}\right) \log\left(\mathcal{N}_p(\nu/2,\mathcal{G},\mathcal{R}^X)\right)} d\nu\right) \\ \inf_{\epsilon\in[0,\mathcal{G}_{\max}(X,Y)/2]} \left(4\epsilon + \frac{12}{\sqrt{|\mathcal{R}^X|}} \int_\epsilon^{\mathcal{G}_{\max}(X,Y)/2} \sqrt{\log\left(\mathcal{N}_\infty(\nu/2,\mathcal{G},\mathcal{R}^X \cup \{(X,Y)\})\right)} d\nu\right). \end{cases}$$

As a corollary we obtain the following rates, as the log-covering number varies with $\nu$ at different rates.

**Corollary B.4.** *Under Assumption 3.1 we have for any retrieved set within radius $r$ of $X$, $\mathcal{R}^X$, for any $p \geq 1$*

$$\mathfrak{R}_{\mathcal{R}^X}\left(\mathcal{G}(X,Y)\right) \leq \begin{cases} C'(p,\mathcal{F}^{\text{loc}}) \frac{\log^2\left(2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}\right)}{\sqrt{|\mathcal{R}^X|}}; & \text{if } \log\left(\mathcal{N}_p(\varepsilon,\mathcal{G},n)\right) \leq C^2(p,\mathcal{F}^{\text{loc}}) \log(n/\varepsilon)/\varepsilon^2, \\ \frac{C'(p,\mathcal{F}^{\text{loc}}) \mathcal{G}_{p,\max}(X,Y)^{1-\alpha/2} \log\left(2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}\right)}{\sqrt{|\mathcal{R}^X|}}; & \text{if } \log\left(\mathcal{N}_p(\varepsilon,\mathcal{G},n)\right) \leq C^2(p,\mathcal{F}^{\text{loc}}) \log(n/\varepsilon)/\varepsilon^\alpha, \alpha \in [0,2). \end{cases}$$

*Proof.* Case $\log\left(\mathcal{N}_p(\varepsilon,\mathcal{G},n)\right) \leq C^2(p,\mathcal{F}^{\text{loc}}) \log(n/\varepsilon)/\epsilon^\alpha$, $\alpha \in [0,2)$:

$$\begin{aligned} \mathfrak{R}_{\mathcal{R}^X}\left(\mathcal{G}(X,Y)\right) &\leq \tfrac{4\mathcal{G}_{p,\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \int_{\mathcal{G}_{p,\max}(X,Y)/\sqrt{|\mathcal{R}^X|}}^{\mathcal{G}_{p,\max}(X,Y)/2} \sqrt{\log\left(\tfrac{2\mathcal{G}_{\max}}{\nu}\right) \log\left(\tfrac{2|\mathcal{R}^X|}{\nu}\right)} \nu^{-\alpha/2} d\nu \\ &\leq \tfrac{4\mathcal{G}_{p,\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \int_{\mathcal{G}_{p,\max}(X,Y)/\sqrt{|\mathcal{R}^X|}}^{\mathcal{G}_{p,\max}(X,Y)/2} \log\left(\tfrac{2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}}{\nu}\right) \nu^{-\alpha/2} d\nu \\ &\leq \tfrac{4\mathcal{G}_{p,\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + C'' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \Big(\mathcal{G}_{p,\max}(X,Y)^{1-\alpha/2} \log\left(2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}\right) + (1-\alpha/2)^{-1} e^{-1}\Big). \end{aligned}$$

The last inequality follows from

$$\begin{aligned} (1-\alpha/2)^2 \int_c^b x^{-\alpha/2} \log(a/x) dx &= b^{1-\alpha/2}(1 + (1-\alpha/2)\log(a/b)) - c^{1-\alpha/2}(1 + (1-\alpha/2)\log(a/b)) \\ &= (b^{1-\alpha/2} - c^{1-\alpha/2})(1 + (1-\alpha/2)\log(a)) + b^{1-\alpha/2}\log(1/b) - c^{1-\alpha/2}\log(1/c) \\ &\leq (b^{1-\alpha/2} - c^{1-\alpha/2})(1 + (1-\alpha/2)\log(a)) + (1-\alpha/2)^{-1} e^{-1} \end{aligned}$$

Case $\log\left(\mathcal{N}_p(\varepsilon,\mathcal{G},n)\right) \leq C^2(p,\mathcal{F}^{\text{loc}}) \log(n/\varepsilon)/\varepsilon^2$:

$$\begin{aligned} \mathfrak{R}_{\mathcal{R}^X}\left(\mathcal{G}(X,Y)\right) &\leq \tfrac{4}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \int_{1/\sqrt{|\mathcal{R}^X|}}^{\mathcal{G}_{p,\max}(X,Y)/2} \sqrt{\log\left(\tfrac{2\mathcal{G}_{\max}}{\nu}\right) \log\left(\tfrac{2|\mathcal{R}^X|}{\nu}\right)} \nu^{-1} d\nu \\ &\leq \tfrac{4}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \int_{1/\sqrt{|\mathcal{R}^X|}}^{\mathcal{G}_{p,\max}(X,Y)/2} \log\left(\tfrac{2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}}{\nu}\right) \nu^{-1} d\nu \\ &\leq \tfrac{4}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \Big(\log^2\left(2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}\sqrt{|\mathcal{R}^X|}\right) - \log^2\left(\tfrac{4\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}}{\mathcal{G}_{p,\max}(X,Y)}\right)\Big) \\ &\leq \tfrac{4}{\sqrt{|\mathcal{R}^X|}} + C' \tfrac{C(p,\mathcal{F}^{\text{loc}})}{\sqrt{|\mathcal{R}^X|}} \log^2\left(2\max\{|\mathcal{R}^X|,\mathcal{G}_{\max}\}\sqrt{|\mathcal{R}^X|}\right). \end{aligned}$$

□

*Proof of Lemma B.3.* Given the set $\mathcal{R}^X$, and some function $g \in \mathcal{G}(X,Y)$ let us define for $p \geq 1$

$$\|g\|_{p,\mathcal{R}^X} = \left(\tfrac{1}{|\mathcal{R}^X|} \sum_{(x',y')\in\mathcal{R}^X} |g(x',y')|^p\right)^{1/p}.$$

Then, we have $\mathcal{G}_{p,\max}\big((X,Y);\mathcal{R}^X\big) = \max_{g\in\mathcal{G}} \|g\|_{p,\mathcal{R}^X}$ for all $g \in \mathcal{G}(X,Y)$. For the sake of brevity we will use $\mathcal{G}_{p,\max}(X,Y)$ in place of $\mathcal{G}_{p,\max}\big((X,Y);\mathcal{R}^X\big)$. Note that we have from previous definition $\mathcal{G}_{\max}(X,Y) = \mathcal{G}_{\infty,\max}(X,Y) \geq \mathcal{G}_{p,\max}(X,Y)$ for any $p \geq 1$.

A simple bound on the Rademacher complexity comes as a function of the radius $r$ but which is independent of the size of $|\mathcal{R}^X|$. Specifically, we have for Rademacher random variable $\sigma_i$-s

$$\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big) \leq \tfrac{1}{|\mathcal{R}^X|}\mathbb{E}_\sigma\Big[\big|\sum_{(X_i',Y_i')\in\mathcal{R}^X}\sigma_i g_i\big|\Big] \leq \max_{g_i\in\mathcal{G}(X,Y)}|g_i| \leq \mathcal{G}_{\max}(X,Y).$$

Next using the Chaining method (Shalev-Shwartz & Ben-David, 2014, Chapter 27) we can bound the Radamacher complexity as

$$\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big) \leq \inf_{\epsilon\in[0,\mathcal{G}_{p,\max}(X,Y)/2]}\Big(4\epsilon + \tfrac{12}{\sqrt{|\mathcal{R}^X|}}\int_\epsilon^{\mathcal{G}_{p,\max}(X,Y)/2}\sqrt{\log\mathcal{N}_p(\nu,\mathcal{G}(X,Y),\mathcal{R}^X)}d\nu\Big).$$

To finish the proof we need to show, for $p \geq 1$

$$\mathcal{N}_p(\nu,\mathcal{G}(X,Y),\mathcal{R}^X) \leq \mathcal{N}_p(\nu/2,\mathcal{G},\mathcal{R}^X)\mathcal{N}_p(\nu/2,\mathcal{G},\{(X,Y)\}).$$

First we fix any $p \geq 1$. Let $\widehat{\mathcal{U}}$ (a set of real numbers) be a $\nu/2$ cover (in $\ell_p$ norm) of $\mathcal{G}$ with respect to $\{(X,Y)\}$. We have $\mathcal{N}_p(\nu,\mathcal{G}(X,Y),\mathcal{R}^X) \leq \frac{2\mathcal{G}_{\max}}{\nu}$ for any $p \geq 1$ and any $\nu > 0$. Further, let $\tilde{\mathcal{U}}$ be a $\nu/2$ cover of $\mathcal{G}$ with respect to $\mathcal{R}^X$. Note for any $\tilde{u} \in \tilde{\mathcal{U}}$ we have $\tilde{u} \in \mathbb{R}^{|\mathcal{R}^X|}$.

Now, we fix any $g' \in \mathcal{G}$. We have at least one $\tilde{u} \in \tilde{\mathcal{U}}$, and $\hat{u} \in \widehat{\mathcal{U}}$ such that

$$\Big(\tfrac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}|g'(x',y') - \tilde{u}(x',y')|^p\Big)^{1/p} \leq \nu/2, \text{ and } |g'(X,Y) - \hat{u}| \leq \nu/2.$$

Therefore,

$$\begin{aligned}
&\Big(\tfrac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}|\big(g'(x',y') - g'(X,Y)\big) - \big(\tilde{u}(x',y') - \hat{u}\big)|^p\Big)^{1/p} \\
&= \Big(\tfrac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}|\big(g'(x',y') - \tilde{u}(x',y')\big) + \big(\hat{u} - g'(X,Y)\big)|^p\Big)^{1/p} \\
&\leq \Big(\tfrac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}|g'(x',y') - \tilde{u}(x',y')|^p\Big)^{1/p} + |\hat{u} - g'(X,Y)| \\
&\leq \nu/2 + \nu/2 \leq \nu
\end{aligned}$$

The first inequality follows by applying Minkowski's inequality. Whereas, for the second inequality we apply Jensen's inequality for $(\cdot)^{1/p}$ being a concave function for $p \geq 1$, and applying the appropriate scaling. Therefore, given the covers $\tilde{\mathcal{U}}$ and $\widehat{U}$, we can construct the set $\mathcal{U}'$ with entries $u' \in \mathbb{R}^{|\mathcal{R}^X|}$ as: $\mathcal{U}' := \{u' = (\tilde{u}(x,y) - \hat{u}) : \tilde{u} \in \tilde{\mathcal{U}}, \hat{u} \in \widehat{\mathcal{U}}\}$. In particular, $|\mathcal{U}'| = |\widehat{\mathcal{U}}||\tilde{\mathcal{U}}|$. As the choice of $g' \in \mathcal{G}$ and $(x',y') \in \mathcal{R}^X$ were arbitrary, we have $\mathcal{U}'$ to be the cover of $\mathcal{G}(X,Y)$.

For $p = \infty$ we can specialize the bound. In particular, consider $\mathcal{U}$ to be a $\nu/2$ cover (in $\ell_\infty$ norm) of $\mathcal{G}$ with respect to $\mathcal{R}^X \cup \{(X,Y)\}$. Then $\mathcal{U}' := \{u' = (\tilde{u}(x,y) - \hat{u}(X,Y)) : \tilde{u} \in \mathcal{U}\}$ creates a (normalized) $\ell_\infty$ cover for $\mathcal{G}$ with respect to $\mathcal{R}^X$. This is true because $\Big(\tfrac{1}{|\mathcal{R}^X|}\sum_{(x',y')\in\mathcal{R}^X}|g'(x',y') - \tilde{u}(x',y')|^p\Big)^{1/p} \leq |g' - \tilde{u}|_\infty = \nu/2$ and $|\hat{u} - g'(X,Y)| \leq |g' - \tilde{u}|_\infty = \nu/2$. This concludes the proof. □

The first term in the above Lemma is similar to the Chaining based Rademacher bounds (Shalev-Shwartz & Ben-David, 2014, Chapter 28) for $\mathcal{G}$, but the $\epsilon$ (in inf and in the integral) varies in $[0,\mathcal{G}_{\max}(X,Y)]$ instead of $[0,\mathcal{G}_{\max}]$. For small $r$ we have $\mathcal{G}_{\max}(X,Y) << \mathcal{G}_{\max}$, which can be leveraged to give tight bounds in certain situations.

**Example: $\mathcal{F}^{\text{loc}} \equiv \ell_\infty$-bounded RKHS (Zhang, 2004):** Let us consider the setting of Zhang (2004). In this setting, given some Reproducing Kernel Hilbert Space (RKHS) $H$, and a function $\tilde{f} \in H$, we can define the function $\tilde{f}(\cdot) = \tilde{f} \circ h_x$ where for some $h \in H$. We further define the set of functions with bounded norm

$$H_A = \{\tilde{f}(\cdot) \in H : \|\tilde{f}\|_H \sup_{x \in \mathcal{X}} \|h_x\|_H \leq A\}.$$

Finally, our local function class can be defined as

$$\mathcal{F}^{\text{loc}} = H_A^{|\mathcal{Y}|} = \{f(\cdot) : f_y(\cdot) \in H_A, \forall y \in \mathcal{Y}\}.$$

We have $\|\mathcal{F}^{\text{loc}}\|_\infty = A$. Recall that loss function for any $y \in \mathcal{Y}$ is given as $\ell(\gamma_f(x, y))$, for any $f \in \mathcal{F}^{\text{loc}}$. We also have for all $y \in \mathcal{Y}$, $|\ell(\gamma_f(x, y)) - \ell(\gamma_{f'}(x, y))| \leq 2L_\ell \sup_y |f_y(x) - f'_y(x)|$ (Zhang, 2004, Assumption 15) with $\gamma_A = 2L_\ell$).

Given the above setting, following Lemma 17 in Zhang (2004) [3], we have for a universal constant $c$

$$\log\Big(\mathcal{N}_\infty(2L_\ell\nu, \mathcal{G}, \mathcal{R}^X \cup \{(X, Y)\})\Big) \leq c|\mathcal{Y}|\|\mathcal{F}^{\text{loc}}\|_\infty^2 \frac{\ln(2 + \|\mathcal{F}^{\text{loc}}\|_\infty/\nu) + \ln(|\mathcal{R}^X| + 1)}{\nu^2}.$$

This gives us the following bound for the Rademacher complexity of $\mathcal{F}^{\text{loc}}$

$$\mathfrak{R}_{\mathcal{R}^X} \leq O\Big(\sqrt{|\mathcal{Y}|}L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty \tfrac{\ln(|\mathcal{R}^X|+1)^{3/2}}{\sqrt{|\mathcal{R}^X|}}\Big). \tag{25}$$

*Proof of Equation* (25). Without optimizing over $\epsilon$ above, we plug in $\epsilon = \frac{\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}}$. We obtain

$$\begin{aligned}
&\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X, Y)\big) \\
&\leq \tfrac{4\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{12}{\sqrt{|\mathcal{R}^X|}} \int_{\frac{\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}}}^{\mathcal{G}_{\max}(X,Y)/2} \sqrt{\log\Big(\mathcal{N}_\infty\Big(\nu/2, \mathcal{G}, \mathcal{R}^X \cup \{(X, Y)\}\Big)\Big)}d\nu \\
&\leq \tfrac{4\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{48\sqrt{c|\mathcal{Y}|}L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty}{\sqrt{|\mathcal{R}^X|}} \int_{\frac{\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}}}^{\mathcal{G}_{\max}(X,Y)/2} \sqrt{\tfrac{\ln(2 + 4L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty/\nu) + \ln(|\mathcal{R}^X| + 1)}{\nu^2}}d\nu \\
&\leq \tfrac{4\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{48\sqrt{c|\mathcal{Y}|}L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty}{\sqrt{|\mathcal{R}^X|}} \int_{\frac{\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}}}^{\mathcal{G}_{\max}(X,Y)/2} \sqrt{\tfrac{\ln((\mathcal{G}_{\max}(X,Y)+4L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty)/\nu)+\ln(|\mathcal{R}^X|+1)}{\nu^2}}d\nu \\
&\leq \tfrac{4\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{48\sqrt{c|\mathcal{Y}|}L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty}{\sqrt{|\mathcal{R}^X|}} \int_{\frac{1}{\sqrt{|\mathcal{R}^X|}}}^{1/2} \sqrt{\tfrac{\ln((1+4L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty/\mathcal{G}_{\max}(X,Y))/\nu')+\ln(|\mathcal{R}^X|+1)}{\nu'^2}}d\nu' \\
&\leq \tfrac{4\mathcal{G}_{\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{32\sqrt{c|\mathcal{Y}|}L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty}{\sqrt{|\mathcal{R}^X|}}\Big(\ln\big((1 + 4L_\ell\|\mathcal{F}^{\text{loc}}\|_\infty/\mathcal{G}_{\max}(X, Y))\sqrt{|\mathcal{R}^X|}\big) + \ln(|\mathcal{R}^X| + 1)\Big)^{3/2}
\end{aligned}$$

We use $\int_x \sqrt{\ln(a/x) + b}/x dx = -2/3(\ln(a/x) + b)^{3/2}$ for the final inequality, and ignore the negative part. □

**Example: $\mathcal{F}^{\text{loc}} \equiv \ell_2$ bounded RKHS (Lei et al., 2019):** We consider a fixed kernel $K(x, x') = \langle\phi(x), \phi(x')\rangle$ for $x, x' \in \mathcal{X}$, and let $H_K$ be the RKHS induced by $K$. Let us define the $\ell_{p,q}$ norm for the vectors $W = (w_1, w_2, \ldots, w_{|\mathcal{Y}|}) \in H_K^{|\mathcal{Y}|}$ as $\|(w_1, \ldots, w_{|\mathcal{Y}|})\|_{p,q} = \|(\|w_1\|_p, \ldots, \|w_{|\mathcal{Y}|}\|_p)\|_q$.

For some norm bound $\Lambda > 0$, the local hypothesis space is defined as

$$\mathcal{F}^{\text{loc}} = \{f(\cdot) : f_y(\cdot) = \langle w_y, \phi(\cdot)\rangle, w_y \in H_K, \forall y \in \mathcal{Y}, \|(w_1, \ldots, w_{|\mathcal{Y}|})\|_{2,2} \leq \Lambda\}.$$

Recall that we have the loss function class $\mathcal{G} = \{\ell(\gamma_f(\cdot, \cdot)) : f \in \mathcal{F}^{\text{loc}}\}$, where the loss function $\ell(\cdot)$ is assumed to be $L$-Lipschitz continuous w.r.t. $\ell_\infty$ norm.

[3] We correct for a typographical error in Zhang (2004), where the $n \equiv |\mathcal{R}^X|$ comes in the denominator of the bound presented in Lemma 17. But Theorem 4 of Zhang (2002) shows this is a typographical error. Indeed, the covering number is not suppossed to decrease with increasing number of points.

Given the retrieved set $\mathcal{R}^X$ for some positive integer $n \geq 1$, $\tilde{\mathcal{F}}^X$ after Equation (8) in Lei et al. (2019) induced by $\mathcal{R}^X$. [4] Let the worst case Rademacher complexity of a function class $\mathcal{F}$ over $n$ points be defined as $\mathfrak{R}_n(\mathcal{F})$. Also, for a set $S$ let $\hat{B}(S) = \max_{(x,y)\in S} \sup_{W:\|W\|_{2,2}\leq\Lambda} \langle w_y, \phi(x)\rangle$. We have from Theorem 23 in Lei et al. (2019) that the covering number is bounded as follows: for any set $S = \{(x_i, y_i) : i = 1, \ldots, n\}$ of size $n \geq 1$, for any $\varepsilon > 4L\mathfrak{R}_{n|\mathcal{Y}|}\big(\tilde{\mathcal{F}}^X\big)$

$$\log\Big(\mathcal{N}_\infty\big(\varepsilon, \mathcal{G}, S\big)\Big) \leq \frac{16n|\mathcal{Y}|L^2(\mathfrak{R}_{n|\mathcal{Y}|}\left(\tilde{\mathcal{F}}^X\right))^2}{\varepsilon^2} \log\big(\tfrac{2en|\mathcal{Y}|\hat{B}(S)L}{\varepsilon}\big).$$

Furthermore, from equation (18) in Lei et al. (2019) we have for any set

$$\tfrac{\Lambda \max_{(x,y)\in S}\|\phi(x)\|_2}{\sqrt{2n|\mathcal{Y}|}} \leq \mathfrak{R}_{n|\mathcal{Y}|}\big(\tilde{\mathcal{F}}^X\big) \leq \tfrac{\Lambda \max_{(x,y)\in S}\|\phi(x)\|_2}{\sqrt{n|\mathcal{Y}|}}.$$

Therefore, we have for all $\varepsilon \geq 4L\frac{\Lambda \max_{(x,y)\in S}\|\phi(x)\|_2}{\sqrt{2n|\mathcal{Y}|}}$

$$\log\Big(\mathcal{N}_\infty\big(\varepsilon, \mathcal{G}, S\big)\Big) \leq \frac{16\max_{(x,y)\in S}\|\phi(x)\|_2^2\Lambda^2L^2}{\varepsilon^2} \log\big(\tfrac{2en|\mathcal{Y}|\hat{B}(S)L}{\varepsilon}\big).$$

Plugging this covering number in in our Rademacher bound with $\epsilon \geq 4L\frac{\Lambda \max_{(x,y)\in S}\|\phi(x)\|_2}{\sqrt{2(|\mathcal{R}^X|+1)|\mathcal{Y}|}}$ and taking $S = \mathcal{R}^X \cup \{(X,Y)\}$ we get

$$\begin{aligned}
\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big) &\leq \inf_{\epsilon\in[0,\mathcal{G}_{\max}(X,Y)/2]}\Big(4\epsilon + \tfrac{12}{\sqrt{|\mathcal{R}^X|}}\int_\epsilon^{\mathcal{G}_{\max}(X,Y)/2}\sqrt{\log\mathcal{N}_\infty(\nu/2, \mathcal{G}, \mathcal{R}^X\cup\{(X,Y)\})}d\nu\Big)\\
&\leq \frac{16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2\Lambda L}{\sqrt{2(|\mathcal{R}^X|+1)|\mathcal{Y}|}} + \frac{12\times16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|\Lambda L}{\sqrt{|\mathcal{R}^X|}}\times\\
&\times\int_{\frac{4L\Lambda\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2}{\sqrt{2(|\mathcal{R}^X|+1)|\mathcal{Y}|}}}^{\mathcal{G}_{\max}(X,Y)/2}\tfrac{1}{\nu}\sqrt{\log\big(\tfrac{4e(|\mathcal{R}^X|+1)|\mathcal{Y}|\hat{B}(\mathcal{R}^X\cup\{(X,Y)\})L}{\nu}\big)}d\nu\\
&\leq \frac{16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2\Lambda L}{\sqrt{2(|\mathcal{R}^X|+1)|\mathcal{Y}|}} + \frac{8\times16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|\Lambda L}{\sqrt{|\mathcal{R}^X|}}\times\\
&\times\Big(\log\big(\tfrac{4\sqrt{2}eL\hat{B}(\mathcal{R}^X\cup\{(X,Y)\})(|\mathcal{R}^X|+1)|\mathcal{Y}|\sqrt{(|\mathcal{R}^X|+1)|\mathcal{Y}|}}{4L\Lambda\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2}\big)\Big)^{3/2}\\
&\leq \frac{16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2\Lambda L}{\sqrt{2(|\mathcal{R}^X|+1)|\mathcal{Y}|}} + \frac{8\times16\max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|\Lambda L}{\sqrt{|\mathcal{R}^X|}}\times\\
&\times\Big(\log\big(\sqrt{2}e\big((|\mathcal{R}^X|+1)|\mathcal{Y}|\big)^{3/2}\big)\Big)^{3/2}
\end{aligned}$$

In the final inequality we use the fact that

$$\begin{aligned}
\hat{B}(\mathcal{R}^X\cup\{(X,Y)\}) &\leq \max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2 \sup_{W:\|W\|_{2,2}\leq\Lambda}\|W\|_{2,\infty}\\
&\leq \max_{(x,y)\in\mathcal{R}^X\cup\{(X,Y)\}}\|\phi(x)\|_2\Lambda
\end{aligned}$$

Therefore, the final bound on the Rademacher complexity can be given as

$$\mathfrak{R}_{\mathcal{R}^X} \leq O\Big(L_\ell\|\mathcal{F}^{\mathrm{loc}}\|_\infty\tfrac{\ln(|\mathcal{Y}||\mathcal{R}^X|)^{3/2}}{\sqrt{|\mathcal{R}^X|}}\Big). \tag{26}$$

**Example: $\mathcal{F}^{\mathbf{loc}} \equiv L$-layer fully connected deep neural network (DNN)(Bartlett et al., 2017):** Following Bartlett et al. (2017), we consider a $L$-layer deep neural network (DNN) $f_{\mathcal{A}} = \sigma_L(A^L\sigma_{L-1}(A^{L-1}\sigma_{L-2}(\ldots A^1x))$ for $x \in \mathcal{X}$ where

[4] We need $\tilde{\mathcal{F}}^X$ only to state some theorems in Lei et al. (2019). We refer interested readers to Lei et al. (2019) for the details.

$\mathcal{A} = (A_1, A_2, \ldots, A_L)$ is the sequence of weight matrices. The matrix $A^l \in \mathbb{R}^{d_{l-1} \times d_l}$ for $l = 1$ to $L$, with $d_L = |\mathcal{Y}|$, and $d_0 = d$ given $\mathcal{X} \subseteq \mathbb{R}^d$. Furthermore, $\sigma_l(\cdot) : \mathbb{R}^{d_l} \to \mathbb{R}^{d_l}$ denotes the non-linearity (including pooling and activation), $\sigma_l$-s are taken to be 1-Lipschitz, and $\sigma_l(0) = 0$. We assume that the $A^l$ matrix is initialized at $M^l$, for each $l = 1$ to $L$. We consider the local function class

$$\mathcal{F}^{\text{loc}} = \{f_{\mathcal{A}} : \|A^l - M^l\|_{2,1} \leq b_l, \|A^l\|_\sigma \leq s_l, \ \forall l \leq l \leq L-1\}.$$

Furthermore, we have for any $f \in \mathcal{F}^{\text{loc}}$ and any $x \in \mathcal{X}$ the function $(f(x), y) \to \ell(\gamma_f(\cdot, \cdot))$ is $2L_\ell$ -Lipschitz. Therefore, for a fixed set $S$, we have from Theorem 3.3 in Bartlett et al. (2017) that the covering number of the $\mathcal{G} = \{\ell(\gamma_f(\cdot, \cdot)) : f_{\mathcal{A}} \in \mathcal{F}^{\text{loc}}\}$ is given as

$$\log\Big(\mathcal{N}_2\big(\varepsilon, \mathcal{G}, S\big)\Big) \leq \frac{4L_\ell^2 B^2 ln(2d_{\max}^2)}{\varepsilon^2}\big(\prod_{l=1}^{L} s_l\big)^2\big(\sum_{l=1}^{L}(b_l/s_l)^{2/3}\big)^{3/2} = \frac{R}{\varepsilon^2},$$

where $d_{\max} = \max_{l=1}^{L} d_l$, $\sqrt{\frac{1}{|S|}\sum_{x\in S}\|x\|_2^2} \leq B$, and

$$R = 4L_\ell^2 B^2 ln(2d_{\max}^2)\big(\prod_{l=1}^{L} s_l\big)^2\big(\sum_{l=1}^{L}(b_l/s_l)^{2/3}\big)^{3/2}.$$

Using a the covering number based bound on Rademacher complexity we obtain

$$\begin{aligned}
&\mathfrak{R}_{\mathcal{R}^X}\big(\mathcal{G}(X,Y)\big) \\
&\leq \inf_{\epsilon\in[0,\mathcal{G}_{2,\max}(X,Y)/2]}\Big(4\epsilon + \tfrac{12}{\sqrt{|\mathcal{R}^X|}}\int_{\epsilon}^{\mathcal{G}_{2,\max}(X,Y)/2}\sqrt{\log(\tfrac{4L_\ell B\prod_{l=1}^{L}s_l}{\nu})\log\Big(\mathcal{N}_2\Big(\nu/2,\mathcal{G},\mathcal{R}^X\Big)\Big)}d\nu\Big) \\
&\leq \inf_{\epsilon\in[0,\mathcal{G}_{2,\max}(X,Y)/2]}\Big(4\epsilon + \tfrac{12}{\sqrt{|\mathcal{R}^X|}}\int_{\epsilon}^{\mathcal{G}_{\max}(X,Y)/2}\sqrt{\log(\tfrac{4L_\ell B\prod_{l=1}^{L}s_l}{\nu})\tfrac{R}{\nu^2}}d\nu\Big) \\
&\leq \inf_{\epsilon\in[0,\mathcal{G}_{2,\max}(X,Y)/2]}\Big(4\epsilon + \tfrac{8\sqrt{R}}{\sqrt{|\mathcal{R}^X|}}\log^{3/2}(\tfrac{4L_\ell B\prod_{l=1}^{L}s_l}{\epsilon})\Big) - \tfrac{8\sqrt{R}}{\sqrt{|\mathcal{R}^X|}}\log^{3/2}(\tfrac{8L_\ell B\prod_{l=1}^{L}s_l}{\mathcal{G}_{2,\max}(X,Y)}) \\
&\leq \Big(\tfrac{4\mathcal{G}_{2,\max}(X,Y)}{\sqrt{|\mathcal{R}^X|}} + \tfrac{8\sqrt{R}}{\sqrt{|\mathcal{R}^X|}}\log^{3/2}(\tfrac{4L_\ell B\prod_{l=1}^{L}s_l\sqrt{|\mathcal{R}^X|}}{\mathcal{G}_{2,\max}(X,Y)})\Big) - \tfrac{8\sqrt{R}}{\sqrt{|\mathcal{R}^X|}}\log^{3/2}(\tfrac{8L_\ell B\prod_{l=1}^{L}s_l}{\mathcal{G}_{2,\max}(X,Y)})
\end{aligned}$$

### B.4. Function approximation

The following proposition states that the expected loss between $\mathcal{F}^{\text{loc}}$ classes can be bounded by the $L_1$ regression error between these two.

**Proposition B.5.** *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be two classes for scorer functions, and loss $\ell$ satisfy Assumption 3.2 with Lipschitz constant $L_\ell$. Then for any $x \in \mathcal{X}$ the following holds*

$$\begin{aligned}
\min_{f\in\mathcal{F}_1}\mathbb{E}_{\mathsf{D}^{x,r}}[\ell(f(X),Y)] \leq \min_{f\in\mathcal{F}_2}\mathbb{E}_{\mathsf{D}^{x,r}}[\ell(f(X),Y)] \\
+ 2L_\ell\max_{f'\in\mathcal{F}_2}\min_{f\in\mathcal{F}_1}\mathbb{E}_{\mathsf{D}^{x,r}}[\max_s|f_s(X) - f'_s(X)|].
\end{aligned}$$

*Proof.* We compare the loss from two arbitrary (measurable w.r.t. $D^{X,r}$) functions $f$ and $\tilde{f}$ next, where we are trying to approximate $\tilde{f}$ with $f$.

$$\begin{aligned}
\mathbb{E}[\ell(f(X),Y)] &= \mathbb{E}[\ell(\gamma_f(X,Y))] \\
&= \mathbb{E}[\ell(\gamma_{\tilde{f}}(X,Y))] + \mathbb{E}[\ell(\gamma_f(X,Y)) - \ell(\gamma_{\tilde{f}}(X,Y))] \\
&\leq \mathbb{E}[\ell(\gamma_{\tilde{f}}(X,Y))] + L_\ell\mathbb{E}[|\gamma_f(X,Y) - \gamma_{\tilde{f}}(X,Y)|] \\
&= \mathbb{E}[\ell(\tilde{f}(X),Y)] + L_\ell\mathbb{E}[|f_Y(X) - \tilde{f}_Y(X) - \max_{s\neq Y}f_s(X) + \max_{s\neq Y}\tilde{f}_s(X)|] \\
&\leq \mathbb{E}[\ell(\tilde{f}(X),Y)] + 2L_\ell\mathbb{E}[\max_s|f_s(X) - \tilde{f}_s(X)|]
\end{aligned}$$

Let $f^{X,*} := \arg\min_{f\in\mathcal{F}^X} \mathbb{E}[\ell(f(X),Y)]$. Replacing the $\tilde{f} = f^{X,*}$ and taking minima over $\mathcal{F}^{\text{loc}}$ on both sides, we obtain.

$$\min_{f\in\mathcal{F}^{\text{loc}}} \mathbb{E}[\ell(f(X),Y)] \le \min_{f\in\mathcal{F}^X} \mathbb{E}[\ell(f(X),Y)] + 2L_\ell \min_{f\in\mathcal{F}^{\text{loc}}} \mathbb{E}[\max_s |f_s(X) - f_s^{X,*}(X)|].$$

□

Applying the above result, we have $\varepsilon_{\text{loc}} = 2L_\ell \min_{f\in\mathcal{F}^{\text{loc}}} \mathbb{E}[\max_s |f_s(X) - f_s^{X,*}(X)|]$ in Eq. (6).

A similar argument establishes that $\varepsilon_X = 2L_\ell \min_{f\in\mathcal{F}^{\text{x}}} \mathbb{E}[\max_s |f_s(X) - f_s^{*}(X)|]$ in Eq. (4), where the function $f^*$ is the population minimizer of the loss over distribution $D^X$ among the function class $\mathcal{F}^{\text{global}}$.

### B.5. Proof of Proposition 3.6

Under the weak density condition, for any $r > 0$ we have $\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^X| = 0\big] = 0$. Furthermore, for any $N \ge 1$, and $x \in \mathcal{X}$

$$\begin{aligned}
&\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^x| < N\big] \le \mathbb{P}_{\mathsf{D}}\big[\sum_{i=1}^n \mathbb{1}\big(d(X_i,x) \le r\big) < N\big] \\
&\le \mathbb{P}_{\mathsf{D}}\big[\sum_{i=1}^n \mathbb{1}\big(d(X_i,x) \le \min\{r, \delta_{\text{wdc}}\rho_{\mathsf{D}}(x)^{-1/d}\}\big) < N\big] \\
&\le \exp(-2(p(x,r) - N/n)^2 n)
\end{aligned}$$

Let $p(x,r) := \min\{c_{\text{wdc}}^d \rho_{\mathsf{D}}(x) r^d, c_{\text{wdc}}^d \delta_{\text{wdc}}^d\}$, then $\mathbb{P}_{\mathsf{D}}\big(d(X_i,x) \le \min\{r, \delta_{\text{wdc}}\rho_{\mathsf{D}}(x)^{-1/d}\}\big) \ge p(x,r)$. Using Chernoff bound we obtian the final inequality for the above definition of $p(x,r)$. It can be shown that choosing

$$N = n\big(\min\{c_{\text{wdc}}^d \rho_{\mathsf{D}}(x) r^d, c_{\text{wdc}}^d \delta_{\text{wdc}}^d\} - \sqrt{\tfrac{\log(1/\delta)}{2n}}\big),$$

we obtain $\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^x| < N\big] \le \delta$ for any $\delta > 0$.

Recall, $\mathbb{P}_{\mathsf{D}}[\rho_{\mathsf{D}}(X) < f_\rho(\delta)] \le \delta\}$ for any $\delta > 0$. Let

$$N(r,\delta) \ge n\big(\min\{c_{\text{wdc}}^d f_\rho(\delta) r^d, c_{\text{wdc}}^d \delta_{\text{wdc}}^d\} - \sqrt{\tfrac{\log(2/\delta)}{2n}}\big).$$

Then, we have

$$\begin{aligned}
&\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^X| < N(r,\delta)\big] \\
&\le \mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^X| < N(r,\delta) | \rho_{\mathsf{D}}(X) \ge f_\rho(\delta/2)\big] + \mathbb{P}_{\mathsf{D}}[\rho_{\mathsf{D}}(X) < f_\rho(\delta/2)] \\
&\le \delta/2 + \delta/2 = \delta
\end{aligned}$$

For the first term in the final inequality, we use the fact that for all $x \in \mathcal{X}$ such that $\rho_{\mathsf{D}}(x) \ge \rho_{\mathsf{D}}^{-1}(\delta/2)$, we have $\mathbb{P}_{\mathsf{D}}\big[|\mathcal{R}^x| < N(r,\delta)\big] \le \delta/2$. For the second term in the final inequality, we just use the definition of $f_\rho(\delta)$.

### B.6. Computation of the function $f_\rho(\delta)$ in Proposition 3.6

For non-degenerate multi-dimensional Gaussian distributions we have

$$\rho_{N(\mu,\Sigma)}(x) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

Therefore, the level sets are given as

$$\begin{aligned}
&\mathbb{P}_{N(\mu,\Sigma)}\big[x : \rho_{N(\mu,\Sigma)}(x) \ge (2\pi)^{-d/2}|\Sigma|^{-1/2}\gamma\big] \\
&= \int_{x:(x-\mu)^T\Sigma^{-1}(x-\mu)\le 2\ln(1/\gamma)} (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))dx \\
&= \int_{x:(x-\mu)^T\Sigma^{-1}(x-\mu)\le 2\ln(1/\gamma)} (2\pi)^{-d/2}|\Sigma|^{-1/2} \int_{t=\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}^{\infty} \exp(-t)dtdx
\end{aligned}$$

$$= (2\pi)^{-d/2}|\Sigma|^{-1/2}\int_{t=0}^{\infty}\underbrace{\Big(\int_{x:(x-\mu)^T\Sigma^{-1}(x-\mu)\leq \min\{2t,2\ln(1/\gamma)\}} dx\Big)}_{\text{volume of ellipsoid}}\exp(-t)dt$$

$$= (2\pi)^{-d/2}|\Sigma|^{-1/2}|\Sigma|^{1/2}\tfrac{\pi^{d/2}}{\Gamma(d/2+1)}\Big(\int_{t=\ln(1/\gamma)}^{\infty}(2\ln(1/\gamma))^{d/2}\exp(-t)dt + \int_{t=0}^{\ln(1/\gamma)}(2t)^{d/2}\exp(-t)dt\Big)$$

$$= \tfrac{1}{\Gamma(d/2+1)}\big(\Gamma(d/2+1) - \int_{\ln(1/\gamma)}^{\infty}(t^{d/2} - \ln(1/\gamma)^{d/2})\exp(-t)dt\big)$$

$$= 1 - \tfrac{1}{\Gamma(d/2+1)}\int_{0}^{\infty}((t'+\ln(1/\gamma))^{d/2} - \ln(1/\gamma)^{d/2})\exp(-(t'+\ln(1/\gamma)))dt'$$

$$\geq 1 - \tfrac{\gamma}{\Gamma(d/2+1)}\int_{0}^{\infty}(q(t'/q)^{d/2} + r(\ln(1/\gamma)/r)^{d/2} - \ln(1/\gamma)^{d/2})\exp(-t'))dt' \quad [(q\alpha+r\beta)^p \leq q\alpha^p + r\beta^p : q+r=1]$$

$$\geq 1 - \Big(\ln(1/\gamma)^{d/2-1}\gamma + \tfrac{((1-1/ln(1/\gamma))^{-d/2+1}-1)}{\Gamma(d/2+1)}\gamma\ln(1/\gamma)^{d/2}\Big)$$

$$\geq 1 - \Big(\ln(1/\gamma)^{d/2-1}\gamma + \tfrac{((1-1/ln(1/\gamma))^{-d/2+1}-1)}{\Gamma(d/2+1)}\gamma\ln(1/\gamma)^{d/2}\Big) \geq 1 - 2.45\gamma\ln(1/\gamma)^{d/2},\ \forall\gamma\leq 1/2$$

We now extend the results to mixture of distributions. It is easily shown below that if each of the mixture component $k \leq K$ satisfies $\mathbb{P}_{\mathsf{D}_k}\big[x : \rho_{\mathsf{D}_k}(x) \leq \gamma\big] \leq c\gamma\ln(1/\gamma)^{d/2}$ then

$$\mathbb{P}_{\mathsf{D}_{\text{mix}}}\big[x : \rho_{\mathsf{D}_{\text{mix}}}(x) \leq \gamma\big] = \sum_k w_k\mathbb{P}_{\mathsf{D}_k}\big[x : \sum_l w_l\rho_{\mathsf{D}_l}(x) \leq \gamma\big] \leq \sum_k w_k\mathbb{P}_{\mathsf{D}_k}\big[x : \rho_{\mathsf{D}_k}(x) \leq \gamma w_k^{-1}\big]$$
$$\leq \sum_k w_k c\gamma w_k^{-1}\ln(w_k/\gamma)^{d/2} \leq cK\gamma\ln(1/\gamma)^{d/2}.$$

## C. Comparison of risk bounds

We now compare the risk bounds of the proposed explicit local ERM (which we think of as proxy towards understanding the implicit local learning happening in Retrieval augmented models) between different parametric, and non-parametric methods.

### C.1. Sobolev spaces (Yarotsky, 2017)

In the following paragraph we briefly describe the setting of (Yarotsky, 2017) for the most part borrowing the notations from the authors. The authors study the approximation of functions $f : \mathbb{R}^d \to \mathbb{R}$, with Relu networks with the metrics $\max_{x\in[0,1]^d}|f(x) - \tilde{f}(x)|$ for some approximation $\tilde{f}$. They consider the Sobolev spaces $\mathcal{W}^{k,\infty}([0,1]^d)$ for $n = 1, 2, \ldots$. For a function in Sobolev spaces $\mathcal{W}^{k,\infty}([0,1]^d)$ the weak derivatives upto order $k$ are bounded in $L_\infty$ norm. In particular, we define the norm in $\mathcal{W}^{k,\infty}([0,1]^d)$ as .

$$\|f\|_{\mathcal{W}^{k,\infty}([0,1]^d)} = \max_{|\mathbf{k}|\leq k}\operatorname*{ess\,sup}_{x\in[0,1]^d}\|D^{\mathbf{k}}f(x)\|_\infty,$$

where $\mathbf{k} \in \{0, 2, \ldots, k\}^d$ is the multi-index of the weak derivative $D^{\mathbf{k}}$, and $|\mathbf{k}| = \sum_{k_i=1}^d k_i$. [5] The function class

$$F_{k,d} = \{f \in \mathcal{W}^{k,\infty}([0,1]^d) : \|f\|_{\mathcal{W}^{k,\infty}([0,1]^d)} \leq 1\}.$$

From Theorem 4 we know that to approximate $F_{k,d}$ within accuracy $\epsilon \in (0, 1/2)$ we require $\Omega(\epsilon^{-d/2k})$ weights in general, and with a depth $O(\ln^p(1/\epsilon))$ network, for any $p \geq 0$, we require $\Omega(\epsilon^{-d/k}\ln^{-(2p+1)}(1/\epsilon))$ weights.

A standard bound of Taylor series ensures that a degree $(k-1)$ Taylor polynomial will approximate the function class $F_{k,d}$ for any $x'$ in the $L_2$-radius $r$ of $x$ (hence $L_\infty$-radius $r$ as $L_2$ norm upper bounds $L_\infty$ norm) with accuracy $d^k r^k/k!$, i.e. for any $f \in F_{k,d}$ there exists $\tilde{f}(x') \in \mathcal{P}(k)$

$$\max_{x':\|x'-x\|_2\leq r}|f(x') - \tilde{f}(x')| \leq \tfrac{d^k r^k}{k!}\|f\|_{\mathcal{W}^{k,\infty}([0,1]^d)} \leq \tfrac{d^k r^k}{k!}.$$

[5] Note we adopt bold symbol only for multi-indices, whereas vectors in $\mathbb{R}^d$ are denoted without bold symbol.

In particular, $\tilde{f}(x')$ can be taken as the Taylor polynomial of degree at most $k$ $\tilde{f}(x') = \sum_{|\mathbf{k}|\le k-1} a_{\mathbf{k}}(x-x')^{\mathbf{k}}$ where $|a_{\mathbf{k}}| \le 1$. Hence, we have $\sum_{|\mathbf{k}|\le k-1} a_{\mathbf{k}} \le \binom{d+k-1}{k-1}$.

Connecting back to our approximation with the approximation error $\varepsilon_{\text{loc}}$ we have for $\mathcal{F}^{\text{global}} = F_{k,d}$ the degree $q_x = (k-1)$ for all $x \in \mathcal{X}$, $C(F_{k,d}, k-1) = \frac{d^k r^k}{k!}$, and $C'(F_{k,d}, k-1) = \binom{d+k-1}{k-1}$.

## D. Proofs for Section 3.4

This section focuses on providing a proof of Proposition 3.8. It follows the proof technique of (Foster et al., 2019, Eq. (9)). Before presenting the proof of Proposition 3.8, we need to introduce a slight variation of the Rademacher complexity for data-dependent hypothesis set.

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\mathcal{R} = \{z_j^{\mathcal{R}}\}, \mathcal{T} = \{z_j^{\mathcal{T}}\} \in \mathcal{Z}^m$ be two $m$-sized samples and $\boldsymbol{\sigma} \in \{+1,-1\}^m$ be a vector of independent Rademacher variables. Now define $\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}} = \{z_j^{\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}}\} \in \mathcal{Z}^m$ such that

$$z_j^{\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}} = \begin{cases} z_j^{\mathcal{R}}, & \text{if } \sigma_j = 1, \\ z_j^{\mathcal{T}}, & \text{if } \sigma_j = -1, \end{cases} \tag{27}$$

i.e., $\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}$ is obtained by replacing $i$-th element of $\mathcal{R}$ by $i$-th element of $\mathcal{T}$ iff $\sigma_i = -1$. Let $\mathcal{U} \in \mathcal{Z}^{n-m}$ be an $m-n$-sized sample; for $\mathcal{R} \in \mathcal{Z}^m$, $\mathcal{S}_{\mathcal{R}} = \mathcal{U} \cup \mathcal{R} \in \mathcal{Z}^n$. Note that, following this notation, we have $\mathcal{S}_{\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}} = \mathcal{U} \cup \mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}$. For $\mathcal{S} \in \mathcal{Z}^n$, let $\mathcal{H}(\mathcal{S})$ be a data dependent function class (hypothesis set), which does not depend on the ordering of the elements in $\mathcal{S}$.

**Definition D.1** (Rademacher complexity for data-dependent function class)**.** *Let $\mathcal{H} = \{\mathcal{H}(\mathcal{S})\}_{\mathcal{S}\in\mathcal{Z}^n}$ be a family of data dependent function classes. Given $\mathcal{R} = \{z_{j\in[m]}^{\mathcal{R}}\}, \mathcal{T} = \{z_{j\in[m]}^{\mathcal{T}}\} \sim \mathsf{D}^m$ and $\mathcal{U} = \{z_{m+i}^{\mathcal{U}}\}_{i\in[n-m]}$, the empirical Rademacher complexity $\mathfrak{R}^{\diamond}_{\mathcal{U},\mathcal{R},\mathcal{T}}(\mathcal{H})$ and Rademacher complexity $\mathfrak{R}^{\diamond}_{\mathcal{U},m}(\mathcal{H})$ are defined as follows.*

$$\begin{aligned} \mathfrak{R}^{\diamond}_{\mathcal{U},\mathcal{R},\mathcal{T}}(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}(\mathcal{S}_{\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}})} \sum_{i=1}^m \sigma_i h(z_i^{\mathcal{T}})\right] \\ \mathfrak{R}^{\diamond}_{\mathcal{U}}(\mathcal{H}) &= \frac{1}{m}\mathbb{E}_{\mathcal{R},\mathcal{T}\sim\mathsf{D}^m,\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}(\mathcal{S}_{\mathcal{R}_{\mathcal{T},\boldsymbol{\sigma}}})} \sum_{i=1}^m \sigma_i h(z_i^{\mathcal{T}})\right] \end{aligned} \tag{28}$$

### D.1. Proof of Proposition 3.8

We are now ready to establish the proof of Proposition 3.8. As discussed above, we extend the proof technique of (Foster et al., 2019, Eq. (9)) to obtain this result. Our setting differs from that of (Foster et al., 2019) as the local ERM objective only depends on the retrieve samples $\mathcal{R}^x$ while the function class of interest $\mathcal{F}_{\mathcal{S}} = \mathcal{F}_{\Phi_{\mathcal{S}}}$ in (16) depends on the entire training set $\mathcal{S}$ via representation $\Phi_{\mathcal{S}}$. We suitably modify the proof techniques of (Foster et al., 2019) to handle this difference.

Let $|\mathcal{R}^x| := m$ and $\mathcal{U} = \mathcal{S}\backslash\mathcal{R}^x$. For $\mathcal{R}, \mathcal{T} \in \mathcal{Z}^m$, we define

$$\begin{aligned} \Xi(\mathcal{R},\mathcal{T}) &= \sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}} \left| \underbrace{\mathbb{E}_{(X',Y')\sim\mathsf{D}^{x,r}}[\ell(f(X'),Y')]}_{:=R_\ell(f;\mathsf{D}^{x,r})} - \underbrace{\frac{1}{m}\sum_{(x',y')\in\mathcal{T}} \ell(f(x'),y')}_{:=\widehat{R}_\ell(f;\mathcal{T})} \right| \\ &= \sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}} \left| R_\ell(f;\mathsf{D}^{x,r}) - \widehat{R}_\ell(f;\mathcal{T}) \right|. \end{aligned}$$

Note that we are interested in bounding

$$\Xi(\mathcal{R}^x,\mathcal{R}^x) = \sup_{f\in\mathcal{F}_{\Phi_{\mathcal{S}}}} \left| \underbrace{\mathbb{E}_{(X',Y')\sim\mathsf{D}^{x,r}}[\ell(f(X'),Y')]}_{R_\ell(f;\mathsf{D}^{x,r})} - \underbrace{\frac{1}{m}\sum_{(x',y')\in\mathcal{T}} \ell(f(x'),y')}_{\widehat{R}_\ell(f;\mathcal{R}^x)=\widehat{R}^x_\ell(f)} \right|,$$

where we have used the fact that $\mathcal{U}\cup\mathcal{R}^x = \mathcal{S}$. Towards this, we first establish that $\Xi(\mathcal{R},\mathcal{R})$ satisfies the $\left(\frac{M_\ell}{m} + 2\Delta L L_{\ell,1}\right)$-bounded difference property, i.e., for $\mathcal{R}, \mathcal{R}' \in \mathcal{Z}^m$ that only differ in one element, we have

$$\Xi(\mathcal{R},\mathcal{R}) - \Xi(\mathcal{R}',\mathcal{R}') \le \frac{M_\ell}{m} + 2\Delta L L_{\ell,1}. \tag{29}$$

Note that

$$\Xi(\mathcal{R},\mathcal{R})-\Xi(\mathcal{R}',\mathcal{R}')\leq \underbrace{\Xi(\mathcal{R},\mathcal{R})-\Xi(\mathcal{R},\mathcal{R}')}_{\mathrm{I}}+\underbrace{\Xi(\mathcal{R},\mathcal{R}')-\Xi(\mathcal{R}',\mathcal{R}')}_{\mathrm{II}}. \tag{30}$$

Now, we will separately bound the two terms in the RHS. Let $\breve{z}=(\breve{x},\breve{y})\in\mathcal{R}\backslash\mathcal{R}'$ and $\breve{z}'=(\breve{x}',\breve{y}')\in\mathcal{R}'\backslash\mathcal{R}$. Thus, we have the following bound on the first term.

$$\begin{aligned}
\mathrm{I}&=\Xi(\mathcal{R},\mathcal{R})-\Xi(\mathcal{R},\mathcal{R}')\\
&=\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R})\right|-\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|\\
&\leq\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\Big|\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R})\right|-\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|\Big|\\
&\leq\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left[R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R})-R_\ell(f;\mathsf{D}^{x,r})+\widehat{R}_\ell(f;\mathcal{R}')\right]\\
&=\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left|\widehat{R}_\ell(f;\mathcal{R}')-\widehat{R}_\ell(f;\mathcal{R})\right|\\
&=\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\frac{1}{m}\left|\ell(f(\breve{x}'),\breve{y}')-\ell(f(\breve{x}),\breve{y})\right|\leq\frac{M_\ell}{m},
\end{aligned} \tag{31}$$

where the last inequality follows from our boundedness assumption for the loss function $\ell$.

Now we move to term II. Towards this, note that, it follows from the definition of supremum that, for any $\epsilon>0$, there exists $\tilde{f}\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}$ such that

$$\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|-\epsilon\leq\left|R_\ell(\tilde{f};\mathsf{D}^{x,r})-\widehat{R}_\ell(\tilde{f};\mathcal{R}')\right| \tag{32}$$

Let $\tilde{f}=\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}}\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}$ and $\tilde{f}'=\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}'}\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}'}}$. Note that, for any $(x,y)\in\mathcal{Z}$,

$$\begin{aligned}
\left|\ell\big(\tilde{f}(x),y\big)-\ell\big(\tilde{f}'(x),y\big)\right|&=\left|\ell\big(\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}}(x),y\big)-\ell\big(\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}'}(x),y\big)\right|\\
&\overset{(i)}{\leq}L_{\ell,1}\|\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}}(x)-\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}'}(x)\|_\infty\\
&\leq L_{\ell,1}\|\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}}(x)-\tilde{g}\circ\Phi_{\mathcal{U}\cup\mathcal{R}'}(x)\|_2\\
&\overset{(ii)}{\leq}L_{\ell,1}L\|\Phi_{\mathcal{U}\cup\mathcal{R}}(x)-\Phi_{\mathcal{U}\cup\mathcal{R}'}(x)\|_2\\
&\overset{(iii)}{\leq}L_{\ell,1}L\Delta,
\end{aligned} \tag{33}$$

where we use $L_{\ell,1}$-Lipschitzness of $\ell$ w.r.t. $\|\cdot\|_\infty$ norm, $L$-Lipschitzness of $g$, and $\Delta$-sensitivity of the representation $\Phi$ in $(i)$, $(ii)$, and $(iii)$, respectively.

Now, we have

$$\begin{aligned}
\mathrm{II}&=\Xi(\mathcal{R},\mathcal{R}')-\Xi(\mathcal{R}',\mathcal{R}')\\
&=\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|-\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}'}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|\\
&\overset{(i)}{\leq}\left|R_\ell(\tilde{f};\mathsf{D}^{x,r})-\widehat{R}_\ell(\tilde{f};\mathcal{R}')\right|+\epsilon-\sup_{f\in\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}'}}}\left|R_\ell(f;\mathsf{D}^{x,r})-\widehat{R}_\ell(f;\mathcal{R}')\right|\\
&\leq\left|R_\ell(\tilde{f};\mathsf{D}^{x,r})-\widehat{R}_\ell(\tilde{f};\mathcal{R}')\right|+\epsilon-\left|R_\ell(\tilde{f}';\mathsf{D}^{x,r})-\widehat{R}_\ell(\tilde{f}';\mathcal{R}')\right|\\
&=\Big|\left[R_\ell(\tilde{f};\mathsf{D}^{x,r})-R_\ell(\tilde{f}';\mathsf{D}^{x,r})\right]-\left[\widehat{R}_\ell(\tilde{f};\mathcal{R}')-\widehat{R}_\ell(\tilde{f}';\mathcal{R}')\right]\Big|+\epsilon\\
&\leq\left|R_\ell(\tilde{f};\mathsf{D}^{x,r})-R_\ell(\tilde{f}';\mathsf{D}^{x,r})\right|+\left|\widehat{R}_\ell(\tilde{f};\mathcal{R}')-\widehat{R}_\ell(\tilde{f}';\mathcal{R}')\right|+\epsilon\\
&\overset{(ii)}{\leq}2L_{\ell,1}L\Delta+\epsilon,
\end{aligned} \tag{34}$$

where $(i)$ and $(ii)$ follow from (32) and (33), respectively. Now, since $\epsilon$ in (32) can be chosen arbitrarily small, it follows from (30), (31), and (34) that

$$\Xi(\mathcal{R}, \mathcal{R}) - \Xi(\mathcal{R}', \mathcal{R}') \leq \frac{M_\ell}{m} + 2\Delta L L_{\ell,1},$$

i.e., $\Xi(\mathcal{R}, \mathcal{R})$ indeed satisfies the $\big(\frac{M_\ell}{m} + 2\Delta L L_{\ell,1}\big)$-bounded difference property. Now, it follows from the McDiarmid's inequality that, for $\delta > 0$, we have with probability at least $1 - \delta$:

$$\Xi(\mathcal{R}^x, \mathcal{R}^x) \leq \mathbb{E}\big[\Xi(\mathcal{R}^x, \mathcal{R}^x)\big] + \big(M_\ell + 2\Delta L L_{\ell,1} m\big)\sqrt{\frac{\log(1/\delta)}{2m}}$$

or

$$\sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}}}} \big|R_\ell(f; \mathsf{D}^{x,r}) - \widehat{R}^x_\ell(f)\big| \leq \mathbb{E}_{\mathcal{R}^x}\Big| \sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}}}} \big[R_\ell(f; \mathsf{D}^{x,r}) - \widehat{R}^x_\ell(f)\big]\Big| + \big(M_\ell + 2\Delta L L_{\ell,1} m\big)\sqrt{\frac{\log(1/\delta)}{2m}}. \tag{35}$$

Now, first statement of Proposition 3.8 follows from (35) and the fact that $m = |\mathcal{R}^x|$.

It follows from the proof steps in Foster et al. (2019, Section E.1) that

$$\mathbb{E}_{\mathcal{R}^x}\Big[\sup_{f \in \mathcal{F}_{\Phi_{\mathcal{S}=\mathcal{U}\cup\mathcal{R}^x}}} \big|R_\ell(f; \mathsf{D}^{x,r}) - \widehat{R}^x_\ell(f)\big|\Big] \leq 2\mathfrak{R}^\diamond_{\mathcal{U}}(\ell \circ \mathcal{F}), \tag{36}$$

where $\mathcal{F} = \{\mathcal{F}_{\Phi_{\mathcal{U}\cup\mathcal{R}}}\}_{\mathcal{R}\in\mathcal{Z}^m}$ and $\mathfrak{R}^\diamond_{\mathcal{U}}$ is defined in (28). This completes the proof of Proposition 3.8. □

## E. Classification in extended feature space: A kernel-based approach

As introduced in Sec. 2.3, our objective is to learn a function $f : \mathcal{X} \times (\mathcal{X} \times \mathcal{Y})^\star \to \mathbb{R}^{|\mathcal{Y}|}$. For a given instance $x$, such a function can leverage its neighboring set $\mathcal{R}^x \in (\mathcal{X} \times \mathcal{Y})^\star$ to improve the prediction on $x$. In this work, we restrict ourselves to a sub-family of such retrieval-based methods that first map $\mathcal{R}^x \sim \mathsf{D}^{x,r}$ to $\hat{\mathsf{D}}^{x,r}$ — an empirical estimate of the local distribution $\mathsf{D}^{x,r}$, which is subsequently utilized to make a prediction for $x$. In particular, the scorers of interest are of the form:

$$(x, \mathcal{R}^x) \mapsto f(x, \hat{\mathsf{D}}^{x,r}) = \big(f_1(x, \hat{\mathsf{D}}^{x,r}), \ldots, f_{|\mathcal{Y}|}(x, \hat{\mathsf{D}}^{x,r})\big) \in \mathbb{R}^{|\mathcal{Y}|}, \tag{37}$$

where $f_y(x, \hat{\mathsf{D}}^{x,r})$ denotes the score assigned to the $y$-th class. Thus, assuming that $\Delta_{\mathcal{X}\times\mathcal{Y}}$ denotes the set of distribution over $\mathcal{X} \times \mathcal{Y}$, we restrict to a suitable function class in $\{f : \mathcal{X} \times \Delta_{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|}\}$. Note that, given a surrogate loss $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ and scorer $f$, the empirical risk $\widehat{R}^{\text{ex}}_\ell(f)$ and population risk $R^{\text{ex}}_\ell(f)$ take the following form:

$$\hat{R}^{\text{ex}}_\ell(f) = \frac{1}{n}\sum\nolimits_{i\in[n]} \ell\big(x_i, \hat{\mathsf{D}}^{x_i,r}\big) \quad \text{and} \quad R^{\text{ex}}_\ell(f) = \mathbb{E}_{(X,Y)\sim\mathsf{D}}\left[\ell\big(f(X, \mathsf{D}^{X,r}), Y\big)\right]. \tag{38}$$

Note that that the general framework for learning in the extended feature space $\widetilde{\mathcal{X}} := \mathcal{X} \times \Delta_{\mathcal{X}\times\mathcal{Y}}$ provides a very rich class of functions. In this paper, we focus on a specific form of learning methods in the extended feature space by using the kernel methods. The method as well as its analysis is obtained by adapting the work on utilizing kernel methods for domain generalization (Blanchard et al., 2011; Deshmukh et al., 2019).

### E.1. Kernel-based classification

Before introducing a kernel method for the classification, we need to define a suitable kernel $k : \widetilde{\mathcal{X}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$ on the extended feature space $\widetilde{\mathcal{X}} := \mathcal{X} \times \Delta_{\mathcal{X}\times\mathcal{Y}}$. Towards this, let $k_{\mathcal{Z}}$ be a kernel over $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. Assuming that $H_{k_{\mathcal{Z}}}$ is the reproducing kernel Hilbert space (RKHS) associated with $k_{\mathcal{Z}}$, we can define a kernel mean embedding (Smola et al., 2007) $\Psi : \Delta_{\mathcal{Z}} \to H_{k_{\mathcal{Z}}}$ as follows:

$$\Psi(P) = \int_{\mathcal{Z}} k_{\mathcal{Z}}\big(z, \cdot\big)\, dP. \tag{39}$$

For an empirical distribution $\hat{\mathsf{D}}^{x,r}$ defined by $\mathcal{R}^x$, kernel embedding in (39) takes the following form.

$$\Psi(\hat{\mathsf{D}}^{x,r}) = \frac{1}{|\mathcal{R}^x|}\sum\nolimits_{(x',y')\in\mathcal{R}^x} k_{\mathcal{Z}}\big((x',y'),\cdot\big). \tag{40}$$

Now, using a kernel $k_{\mathcal{X}}$ over $\mathcal{X}$ and a kernel-like function $\kappa$ over $\Psi(\Delta_{\mathcal{Z}})$, we define a desired kernel $k : \widetilde{\mathcal{X}} \times \widetilde{\mathcal{X}} \to \mathbb{R}$ as follows:

$$k\big(\widetilde{X}_1, \widetilde{X}_2\big) = k\big((X_1, \mathsf{D}^{X_1,r}), (X_2, \mathsf{D}^{X_2,r})\big) = k_{\mathcal{X}}(X_1, X_2) \cdot \kappa\big(\Psi(\mathsf{D}^{X_1,r}), \Psi(\mathsf{D}^{X_2,r})\big). \tag{41}$$

Let $H_k$ be the RKHS corresponding to the kernel $k$ in (41), and $\|\cdot\|_{H_k}$ be the norm associated with $H_k$. Equipped with the kernel in (41) and associated $H_k$, for $\lambda > 0$, we propose to learn a scorer $f = (f_1, \ldots, f_{|\mathcal{Y}|}) \in H_k^{|\mathcal{Y}|} := H_k \times \cdots \times H_k$ via the following regularized ERM problem.

$$\hat{f}^{\text{ex}} = \arg\min_{f\in H_k^{|\mathcal{Y}|}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(\tilde{x}_i), y_i\big) + \lambda \cdot \Omega(f), \tag{42}$$

where $\tilde{x}_i = (x_i, \hat{\mathsf{D}}^{x_i,r})$ and $\Omega(f) := \|f\|^2_{H_k^{|\mathcal{Y}|}} := \sum_{y\in\mathcal{Y}} \|f_y\|^2_{H_k}$. It follows from the representer theorem that the solution of (42) takes the form $\hat{f}^{\text{ex}}(\cdot) = \sum_{i\in[n]} \alpha_i k\big((x_i, \hat{\mathsf{D}}^{x_i,r}), \cdot\big)$. One can apply multiclass extensions of SVMs to learn the weights $\{\alpha_i\}$ (Deshmukh et al., 2019). Next, we focus on studying the generalization behavior of the scorer $\hat{f}^{\text{ex}}$ recovered in (42).

### E.2. Generalization bounds for kernel-based classification

Before presenting a generalization bound for kernel-based classification over the extended feature space $\widetilde{\mathcal{X}}$, we state the three key assumptions that are utilized in our analysis.

**Assumption E.1.** *The loss function $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y}$ is $L_{\ell,1}$-Lipschitz w.r.t. the first argument, i.e.,*

$$|\ell(s_1, y) - \ell(s_2, y)| \le L_{\ell,1} \cdot \|s_1 - s_2\|_\infty \qquad \forall s_1, s_2 \in \mathbb{R}^{|\mathcal{Y}|} \text{ and } y \in \mathcal{Y}. \tag{43}$$

*Furthermore, assume that $\sup_{(x,y)} \ell(x, y) := M_\ell \le \infty$.*

**Assumption E.2.** *Kernels $k_{\mathcal{X}}, k_{\mathcal{Z}}$, and $\kappa$ are bounded by $M_{k_{\mathcal{X}}}, M_{k_{\mathcal{Z}}}$, and $M_\kappa$, respectively.*

**Assumption E.3.** *Let $H_{k_{\mathcal{Z}}}$ and $H_\kappa$ be the RKHS associated with $k_{\mathcal{Z}}$ and $\kappa$, respectively. Then, the canonical feature map $\varphi_\kappa : H_{k_{\mathcal{Z}}} \to H_\kappa$ is $\alpha$-Hölder continuous with $\alpha \in (0, 1]$, i.e.,*

$$\|\varphi_\kappa(h_1) - \varphi_\kappa(h_2)\|_{H_\kappa} \le L' \cdot \|h_1 - h_2\|^\alpha_{H_{k_{\mathcal{Z}}}} \ \forall h_1, h_2 \in \{h \in H_{k_{\mathcal{Z}}} \,:\, \|h\|_{H_{k_{\mathcal{Z}}}} \le M_{k_{\mathcal{Z}}}\} \tag{44}$$

The following result states our generalization bound for the kernel-based classification method described in Sec. E.1.

**Theorem E.4.** *Let $0 \le \delta \le 1$ and Assumptions E.1–E.3 hold. Furthermore, let $N(r, \delta)$ be as defined in* (15)*. Then, for any $B > 0$, the following holds with probability at least $1 - 3\delta$*

$$\begin{aligned}
&\sup_{f\in\mathcal{F}_B^k} \big|\widehat{R}^{\text{ex}}_\ell(f) - R^{\text{ex}}_\ell(f)\big| \le 32\sqrt{\log 2} L_{\ell,1} B M_\kappa M_{k_{\mathcal{X}}} n^{-\frac{1}{2}} \left(1 + \log^{\frac{3}{2}} \sqrt{2} n|\mathcal{Y}|\right) \\
&+ L_{\ell,1} L' M_{k_{\mathcal{X}}} B \left(M_{k_{\mathcal{Z}}} \sqrt{\frac{2\log(\frac{n}{\delta})}{N(r, \frac{\delta}{n})}} + M_{k_{\mathcal{Z}}} \sqrt{\frac{1}{N(r, \frac{\delta}{n})}} + \frac{4 M_{k_{\mathcal{Z}}} \log(\frac{n}{\delta})}{3 N(r, \frac{\delta}{n})}\right)^\alpha + M\sqrt{\frac{\log(\frac{1}{\delta})}{2n}},
\end{aligned}$$

*where $\mathcal{F}_B^k = \big\{f = (f_1, \ldots, f_{|\mathcal{Y}|}) \in H_k^{|\mathcal{Y}|} : \Omega(f) \le B^2\big\}$ and $M := M_\ell + L_{\ell,1} B M_{k_{\mathcal{X}}} M_\kappa$.*

Before presenting the proof of Theorem E.4, we state two key results from the literature that are used in our analysis.

**Proposition E.5** ((Steinwart & Christmann, 2008))**.** *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $H$ be a separable Hilbert space, and $M > 0$. Let $\eta_1, \ldots, \eta_m : \Omega \to H$ be $m$ independent $H$-valued random variables satisfying $\|\eta_j\|_\infty \le M$, for all $j \in [m]$. The, for $\delta > 0$, the following holds with probability at least $1 - \delta$.*

$$\Big\|\frac{1}{m}\sum_{j=1}^{m}(\eta_j - \mathbb{E}_P[\eta_j]\Big\|_H \le M\sqrt{\frac{2\log(1/\delta)}{m}} + M\sqrt{\frac{1}{m}} + \frac{4M\log(1/\delta)}{3m}. \tag{45}$$

**Proposition E.6.** *(Deshmukh et al., 2019; Lei et al., 2019) Let $\widetilde{\mathcal{Z}} = \widetilde{\mathcal{X}} \times \mathcal{Y}$ be (extended) input and output space pair and $\tilde{\mathcal{S}} = \{\tilde{z}_1, \ldots, \tilde{z}_n\}$. Let $H_k$ be a RKHS defined on $\widetilde{\mathcal{X}}$, with $k$ being the associated kernel. Let*

$$\mathcal{F}_B^k = \big\{(f_1, \ldots, f_{|\mathcal{Y}|}) \ : \ f_y \in H_k \ \forall y \in \mathcal{Y} \text{ and } \Big(\sum_{y\in\mathcal{Y}} \|f_y\|_{H_k}^p\Big)^{1/p} \le B\big\}$$

*and $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ be a Lipschitz function in its first argument, i.e.,*

$$|\ell(s_1, y) - \ell(s_2, y)| \le L_{\ell,1}\|s_1 - s_2\|_\infty \quad \forall s_1, s_2 \in \mathbb{R}^{|\mathcal{Y}|} \text{ and } y \in \mathcal{Y}.$$

*Then the Rademacher complexity of the induced function class $\ell \circ \mathcal{F}_B^k := \{\ell \circ f : f \in \mathcal{F}_B^k\}$ satisfies*

$$\begin{aligned}\mathfrak{R}_{\tilde{\mathcal{S}}}\big(\ell \circ \mathcal{F}_B^k\big) &:= \mathbb{E}_{\sigma_i}\Big[\sup_{f\in\mathcal{F}_B^k} \frac{1}{n}\sum_{i\in[n]} \sigma_i \ell\big(f(\tilde{x}_i), y_i\big)\Big] \\ &\le 16 L_{\ell,1}\sqrt{\log 2} B \sup_{\tilde{x}\in\tilde{\mathcal{X}}} \sqrt{k(\tilde{x},\tilde{x})} n^{-\frac{1}{2}} |\mathcal{Y}|^{\frac{1}{2}-\frac{1}{\max\{2,p\}}} \Big(1 + \log^{\frac{3}{2}}\sqrt{2} n|\mathcal{Y}|\Big). \end{aligned} \tag{46}$$

*Note that $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ denotes $n$ i.i.d. Rademacher random variable.*

*Proof of Theorem E.4.* Note that

$$\begin{aligned}\sup_{f\in\mathcal{F}_B^k} \big|\widehat{R}_\ell^{\mathrm{ex}}(f) - R_\ell^{\mathrm{ex}}(f)\big| &= \sup_{f\in\mathcal{F}_B^k} \Big|\frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \widehat{\mathsf{D}}^{x_i,r}), y_i\big) - \mathbb{E}_{(X,Y)\sim\mathsf{D}}\big[\ell\big(f(X, \mathsf{D}^{X,r}), Y\big)\big]\Big| \\ &\le \underbrace{\sup_{f\in\mathcal{F}_B^k} \Big|\frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \widehat{\mathsf{D}}^{x_i,r}), y_i\big) - \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big)\Big|}_{\mathrm{I}} + \\ &\qquad \underbrace{\sup_{f\in\mathcal{F}_B^k} \Big|\frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big) - \mathbb{E}_{(X,Y)\sim\mathsf{D}}\big[\ell\big(f(X, \mathsf{D}^{X,r}), Y\big)\big]\Big|}_{\mathrm{II}} \end{aligned} \tag{47}$$

**Bounding the term-I in** (47)**.** Note that

$$\begin{aligned}\mathrm{I} &= \sup_{f\in\mathcal{F}_B^k} \Big|\frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \widehat{\mathsf{D}}^{x_i,r}), y_i\big) - \frac{1}{n}\sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big)\Big| \\ &\le \frac{L_{\ell,1}}{n}\sum_{i\in[n]} \|f(x_i, \widehat{\mathsf{D}}^{x_i,r}) - f(x_i, \mathsf{D}^{x_i,r})\|_\infty \\ &\le \frac{L_{\ell,1}}{n}\sum_{i\in[n]} \max_{y\in\mathcal{Y}} |f_y(x_i, \widehat{\mathsf{D}}^{x_i,r}) - f_y(x_i, \mathsf{D}^{x_i,r})| \\ &\le L_{\ell,1} \cdot \max_{y\in\mathcal{Y}} \max_{i\in[n]} |f_y(x_i, \widehat{\mathsf{D}}^{x_i,r}) - f_y(x_i, \mathsf{D}^{x_i,r})| \end{aligned} \tag{48}$$

It follows from the reproducing property of the kernel $k$ that, for any $y \in \mathcal{Y}$,

$$\begin{aligned}|f_y(x_i, \widehat{\mathsf{D}}^{x_i,r}) - f_y(x_i, \mathsf{D}^{x_i,r})| &= |\langle f_y, k((x_i, \widehat{\mathsf{D}}^{x_i,r}), \cdot) - k((x_i, \mathsf{D}^{x_i,r}), \cdot)\rangle| \\ &\le \|f_y\|_{H_k} \cdot \|k((x_i, \widehat{\mathsf{D}}^{x_i,r}), \cdot) - k((x_i, \mathsf{D}^{x_i,r}), \cdot)\|_{H_k}. \end{aligned} \tag{49}$$

Now,

$$\begin{aligned}&\|k((x_i, \widehat{\mathsf{D}}^{x_i,r}), \cdot) - k((x_i, \mathsf{D}^{x_i,r}), \cdot)\|_{H_k} \\ &= \Big(k((x_i, \widehat{\mathsf{D}}^{x_i,r}), (x_i, \widehat{\mathsf{D}}^{x_i,r})) + k((x_i, \mathsf{D}^{x_i,r})), (x_i, \mathsf{D}^{x_i,r})) - 2k((x_i, \widehat{\mathsf{D}}^{x_i,r}), (x_i, \mathsf{D}^{x_i,r}))\|_{H_k}\Big)^{1/2}\end{aligned}$$

$$= \sqrt{k_{\mathcal{X}}(x_i,x_i)}\Big(\kappa(\Psi(\widehat{\mathsf{D}}^{x_i,r}),\Psi(\widehat{\mathsf{D}}^{x_i,r}))+\kappa(\Psi(\mathsf{D}^{x_i,r})),\Psi(\mathsf{D}^{x_i,r}))- 2\kappa(\Psi(\widehat{\mathsf{D}}^{x_i,r}),\Psi(\mathsf{D}^{x_i,r}))\|_{H_k}\Big)^{1/2}$$

$$= \sqrt{k_{\mathcal{X}}(x_i,x_i)}\|\kappa(\Psi(\widehat{\mathsf{D}}^{x_i,r}),\cdot)-\kappa(\Psi(\mathsf{D}^{x_i,r}),\cdot)\|_{H_\kappa}$$

$$\leq M_{k_{\mathcal{X}}}\|\kappa(\Psi(\widehat{\mathsf{D}}^{x_i,r}),\cdot)-\kappa(\Psi(\mathsf{D}^{x_i,r}),\cdot)\|_{H_\kappa} \tag{50}$$

$$= M_{k_{\mathcal{X}}}\|\varphi_\kappa(\Psi(\widehat{\mathsf{D}}^{x_i,r}))-\varphi_\kappa(\Psi(\mathsf{D}^{x_i,r}))\|_{H_\kappa}$$

$$\leq L'M_{k_{\mathcal{X}}}\cdot\|\Psi(\widehat{\mathsf{D}}^{x_i,r})-\Psi(\mathsf{D}^{x_i,r})\|^{\alpha}_{H_{k_{\mathcal{Z}}}} \tag{51}$$

By combining (49) and (50), we obtain that

$$|f_y(x_i,\widehat{\mathsf{D}}^{x_i,r})-f_y(x_i,\mathsf{D}^{x_i,r})|\leq L'M_{k_{\mathcal{X}}}\cdot\|f_y\|_{H_k}\cdot\|\Psi(\widehat{\mathsf{D}}^{x_i,r})-\Psi(\mathsf{D}^{x_i,r})\|^{\alpha}_{H_{k_{\mathcal{Z}}}}. \tag{52}$$

Now, Hoeffding's inequality in Hilbert spaces (cf. Proposition E.5) implies that, for $i\in[n]$, the following holds with probability at least $1-\delta$.

$$\|\Psi(\widehat{\mathsf{D}}^{x_i,r})-\Psi(\mathsf{D}^{x_i,r})\|^{\alpha}_{H_{k_{\mathcal{Z}}}} = \Big\|\frac{1}{|\mathcal{R}^{x_i}|}\sum_{(x',y')\in\mathcal{R}^{x_i}}k_{\mathcal{Z}}((x',y'),\cdot)-\mathbb{E}_{\mathsf{D}^{x_i,r}}\big[k_{\mathcal{Z}}((X',Y'),\cdot)\big]\Big\|_{H_{k_{\mathcal{Z}}}}$$
$$\leq M_{k_{\mathcal{Z}}}\sqrt{\frac{2\log(1/\delta)}{|\mathcal{R}^{x_i}|}}+M_{k_{\mathcal{Z}}}\sqrt{\frac{1}{|\mathcal{R}^{x_i}|}}+\frac{4M_{k_{\mathcal{Z}}}\log(1/\delta)}{3|\mathcal{R}^{x_i}|}. \tag{53}$$

It follows from (52) and (53) that, for each $i\in[n]$,

$$|f_y(x_i,\widehat{\mathsf{D}}^{x_i,r})-f_y(x_i,\mathsf{D}^{x_i,r})|$$
$$\leq L'M_{k_{\mathcal{X}}}\cdot\|f_y\|_{H_k}\cdot\Big(M_{k_{\mathcal{Z}}}\sqrt{\frac{2\log(\frac{1}{\delta})}{|\mathcal{R}^{x_i}|}}+M_{k_{\mathcal{Z}}}\sqrt{\frac{1}{|\mathcal{R}^{x_i}|}}+\frac{4M_{k_{\mathcal{Z}}}\log(\frac{1}{\delta})}{3|\mathcal{R}^{x_i}|}\Big)^{\alpha}\quad\forall\, y\in\mathcal{Y} \tag{54}$$

holds with probability at least $1-\delta$. Next, taking union bound over $i\in[n]$ implies that the following holds for all $i\in[n]$ and $y\in\mathcal{Y}$ with probability at least $1-\delta$.

$$|f_y(x_i,\widehat{\mathsf{D}}^{x_i,r})-f_y(x_i,\mathsf{D}^{x_i,r})|$$
$$\leq L'M_{k_{\mathcal{X}}}\|f_y\|_{H_k}\left(M_{k_{\mathcal{Z}}}\sqrt{\frac{2\log(n/\delta)}{|\mathcal{R}^{x_i}|}}+M_{k_{\mathcal{Z}}}\sqrt{\frac{1}{|\mathcal{R}^{x_i}|}}+\frac{4M_{k_{\mathcal{Z}}}\log(n/\delta)}{3|\mathcal{R}^{x_i}|}\right)^{\alpha}. \tag{55}$$

Recall that, for each $i\in[n]$, we have $|\mathcal{R}^{x_i}|\geq N(r,\delta)$ with probability at least $1-\delta$ (cf. (15)). Using union bound, we have $|\mathcal{R}^{x_i}|\geq N(r,\delta/n),\ \forall\, i\in[n]$, with probability at least $1-\delta$. Thus, the following holds for all $i\in[n]$ and $y\in\mathcal{Y}$ with probability at least $1-2\delta$

$$|f_y(x_i,\widehat{\mathsf{D}}^{x_i,r})-f_y(x_i,\mathsf{D}^{x_i,r})|$$
$$\leq L'M_{k_{\mathcal{X}}}\|f_y\|_{H_k}\left(M_{k_{\mathcal{Z}}}\sqrt{\frac{2\log(n/\delta)}{N(r,\delta/n)}}+M_{k_{\mathcal{Z}}}\sqrt{\frac{1}{N(r,\delta/n)}}+\frac{4M_{k_{\mathcal{Z}}}\log(n/\delta)}{3N(r,\delta/n)}\right)^{\alpha}. \tag{56}$$

By using $\|f_y\|_{H_k}\leq B$ and combining (48) with (56), we obtain that

$$\mathrm{I}\leq L_{\ell,1}L'M_{k_{\mathcal{X}}}B\left(M_{k_{\mathcal{Z}}}\sqrt{\frac{2\log(n/\delta)}{N(r,\delta/n)}}+M_{k_{\mathcal{Z}}}\sqrt{\frac{1}{N(r,\delta/n)}}+\frac{4M_{k_{\mathcal{Z}}}\log(n/\delta)}{3N(r,\delta/n)}\right)^{\alpha} \tag{57}$$

holds with probability at least $1-2\delta$.

**Bounding the term-II in** (47)**.** Note that

$$\mathrm{II}=\sup_{f\in\mathcal{F}_B^k}\left|\frac{1}{n}\sum_{i=1}^{n}\ell\big(f(x_i,\mathsf{D}^{x_i,r}),y_i\big)-\mathbb{E}_{(X,Y)\sim\mathsf{D}}\left[\ell\big(f(X,\mathsf{D}^{X,r}),Y\big)\right]\right| \tag{58}$$

Using the Assumptions E.1 and E.2 and the fact that $f \in \mathcal{F}_B^k$, we can argue that

$$\begin{aligned}
\ell\big(f(x, \mathsf{D}^{x,r}), y\big) &= \ell(0, y) + |\ell\big(f(x, \mathsf{D}^{x,r}), y\big) - \ell(0, y)| \\
&\leq M_\ell + L_{\ell,1} \|f(x, \mathsf{D}^{x,r})\|_\infty \\
&\leq M_\ell + L_{\ell,1} \max_{y' \in \mathcal{Y}} \big|\langle f_{y'}, k\big((x, \mathsf{D}^{x,r}), \cdot\big)\rangle\big| \\
&\leq M_\ell + L_{\ell,1} \max_{y' \in \mathcal{Y}} \|f_{y'}\|_{H_k} M_k \\
&\leq M_\ell + L_{\ell,1} R M_k \leq M_\ell + L_{\ell,1} R M_{k_\mathcal{X}} M_\kappa := M
\end{aligned}$$

Now, it follows from the Azuma-McDiarmid's inequality that the following holds with probability at least $1 - \delta$.

$$\begin{aligned}
&\sup_{f \in \mathcal{F}_B^k} \left| \frac{1}{n} \sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big) - \mathbb{E}_{(X,Y)\sim \mathsf{D}} \left[\ell\big(f(X, \mathsf{D}^{X,r}), Y\big)\right] \right| \\
&\qquad \leq \mathbb{E}\left[ \sup_{f \in \mathcal{F}_B^k} \left| \frac{1}{n} \sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big) - \mathbb{E}_{(X,Y)\sim \mathsf{D}} \Big[\ell\big(f(X, \mathsf{D}^{X,r})Y\big)\Big] \right| \right] \\
&\qquad\quad + M\sqrt{\frac{\log(1/\delta)}{2n}},
\end{aligned} \tag{59}$$

Using the standard symmetrization procedure, we get that

$$\begin{aligned}
&\mathbb{E}\left[ \sup_{f \in \mathcal{F}_B^k} \left| \frac{1}{n} \sum_{i=1}^n \ell\big(f(x_i, \mathsf{D}^{x_i,r}), y_i\big) - \mathbb{E}_{(X,Y)\sim \mathsf{D}} \Big[\ell\big(f(X, \mathsf{D}^{X,r})Y\big)\Big] \right| \right] \\
&\qquad \leq \frac{2}{n} \cdot \mathbb{E}_{(X_i,Y_i)\sim \mathsf{D}} \mathbb{E}_{\sigma_i} \left[ \sum_{i \in [n]} \sigma_i \ell\Big(f(x_i, \mathsf{D}^{x_i,r}), y_i\Big) \right], \\
&\qquad = 2\bar{\mathfrak{R}}_{\tilde{\mathsf{S}}}\big(\ell \circ \mathcal{F}_B^k\big)
\end{aligned}$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ denotes $n$ i.i.d. Rademacher random variables and $\bar{\mathfrak{R}}_{\tilde{\mathsf{S}}}\big(\ell \circ \mathcal{F}_B^k\big)$ denote the Rademarcher complexity of the function class

$$\ell \circ \mathcal{F}_B^k = \Big\{ (x, y, \mathsf{D}^{x,r}) \mapsto \ell\big(f(x, \mathsf{D}^{x,r}), y\big) \ : \ f \in \mathcal{F}_B^k \Big\}.$$

Now, using Proposition E.6 with $p = 2$ and Assumption E.2, we have

$$\begin{aligned}
\bar{\mathfrak{R}}_{\tilde{\mathsf{S}}}\big(\ell \circ \mathcal{F}_B^k\big) &\leq 16 L_{\ell,1} \sqrt{\log 2} B \sup_{\tilde{x} \in \tilde{\mathcal{X}}} \sqrt{k(\tilde{x}, \tilde{x})} n^{-\frac{1}{2}} \left(1 + \log^{\frac{3}{2}} \sqrt{2} n |\mathcal{Y}|\right) \\
&\leq 16 L_{\ell,1} \sqrt{\log 2} B M_\kappa M_{k_\mathcal{X}} n^{-\frac{1}{2}} \left(1 + \log^{\frac{3}{2}} \sqrt{2} n |\mathcal{Y}|\right)
\end{aligned} \tag{60}$$

Now, by combining (58), (59), and (60), we obtain that with probability at least $1 - \delta$

$$\mathrm{II} \leq 32 \sqrt{\log 2} L_{\ell,1} B M_\kappa M_{k_\mathcal{X}} n^{-\frac{1}{2}} \left(1 + \log^{\frac{3}{2}} \sqrt{2} n |\mathcal{Y}|\right) + M\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{61}$$

Finally, combining (47), (57) and (61) completes the proof. □

### E.3. Empirical verification

We run an experiment to empirically verify the kernel based extended feature space-based approach. We design a kernel for extended feature space as in (41). In particular, we use Gaussian-like function for

$$\kappa(\Psi(\mathsf{D}^{x_1,r}), \Psi(\mathsf{D}^{x_2,r})) = \exp(-\|\Psi(\mathsf{D}^{x_1,r}) - \Psi(\mathsf{D}^{x_1,r})\|^2 / 2\sigma_\kappa^2).$$

To empirically estimate the distance between kernel mean embeddings of the two distributions $\|\Psi(\hat{\mathsf{D}}^{x_1,r}) - \Psi(\hat{\mathsf{D}}^{x_1,r})\|^2$ we follow Muandet et al. (2017); Li et al. (2015) as:

$$\begin{aligned}\|\Psi(\hat{\mathsf{D}}^{x_1,r}) - \Psi(\hat{\mathsf{D}}^{x_1,r})\|^2 = &\frac{1}{|\mathcal{R}^{x_1}|^2} \sum_{(x',y')\in\mathcal{R}^{x_1}} \sum_{(x'',y'')\in\mathcal{R}^{x_1}} k_{\mathcal{Z}}\big((x',y'),(x'',y'')\big) \\ &+ \frac{1}{|\mathcal{R}^{x_2}|^2} \sum_{(x',y')\in\mathcal{R}^{x_2}} \sum_{(x'',y'')\in\mathcal{R}^{x_2}} k_{\mathcal{Z}}\big((x',y'),(x'',y'')\big) \\ &- \frac{2}{|\mathcal{R}^{x_1}||\mathcal{R}^{x_2}|} \sum_{(x',y')\in\mathcal{R}^{x_1}} \sum_{(x'',y'')\in\mathcal{R}^{x_2}} k_{\mathcal{Z}}\big((x',y'),(x'',y'')\big)\end{aligned}$$

We took $k_{\mathcal{Z}}\big((x',y'),(x'',y'')\big) = \exp(-\|x'-x''\|/2\sigma_x^2 - \lambda\mathbb{1}\{y' \neq y''\})$, which is basically like a normal L2 distance with labels concatenated as one-hot vectors. Also, $k_{\mathcal{X}}(x_1,x_2) = \exp(-\|x_1-x_2\|^2/2\sigma_x^2)$ with normal L2 distance, with which we finally obtain the overall kernel for the extended feature space as:

$$k\big(\widetilde{X}_1,\widetilde{X}_2\big) = k\big((X_1,\mathsf{D}^{X_1,r}),(X_2,\mathsf{D}^{X_2,r})\big) = k_{\mathcal{X}}(X_1,X_2)\cdot\kappa\big(\Psi(\mathsf{D}^{X_1,r}),\Psi(\mathsf{D}^{X_2,r})\big).$$

For the synthetic dataset from Sec. 5, results are tabulated below:

*Table 1.* Accuracy of kernel classifier over extended feature space kernel as a function of number of retrieved neighbors used to form the extended feature space.

| Neighbors | Kernel machine |
|---|---|
| 2 | $0.776 \pm 0.023$ |
| 5 | $0.769 \pm 0.021$ |
| 10 | $0.777 \pm 0.019$ |
| 20 | $0.835 \pm 0.021$ |
| 50 | $0.819 \pm 0.021$ |
| 100 | $0.792 \pm 0.022$ |
| 200 | $0.585 \pm 0.020$ |

Note that, as the generalization bound suggested, that the model performance only improves up to a specific number of neighbors and starts degrading when further increasing the number of neighbors.

It's also worth highlighting that a (4 layer) Transformer model that directly processes an instance along with the associated retrieved neighboring examples achieves a much higher performance of 0.898 with 10 neighbors. This is consistent with similar observations in the deep learning literature where kernel-based methods are often significantly outperformed by end-to-end neural networks (Bai & Lee, 2019; Chen et al., 2020; Samarin et al., 2020).

# F. Additional details for experiments

## F.1. Synthetic

**Task and data.** We consider the task of binary classification on mixtures using *synthetic data*: In particular, we assume $k = 100$ clusters in a $D = 10$-dimensional space. Each cluster is specified by a mean parameter $\mu_i \in \mathbb{R}^D \sim \mathsf{Uniform}(-10, 10)$ and a classification weight vector $w_i \in \mathbb{R}^d \sim \mathcal{N}(0, \mathbb{I})$ for $i = 1, 2, \cdots, k$. We randomly generate a train set of $n = 10000$ points as follows: To generate a labeled example $(x_j, y_j), j \in [n]$: 1) select a cluster $i$ uniformly at random, and 2) sample $x_j \sim \mathcal{N}(\mu_i, \mathbb{I})$ and its label $y_j = \text{sign}(w_i^T(x_j - \mu_i))$. Additionally, we also generate another set of points as test set using the same procedure.

**Methods.** As baseline, we consider models of various complexity, starting from simple linear classifier, to support vector machines with polynomial kernel (of degree 3) and with radial basis function (RBF) kernel, to a multi-layer perceptron (MLP) of two layers. For retrieval-based models, we consider each of the above method as the local model to fit on retrieved data points via local ERM framework (Sec. 3). Additionally, we also report simple kNN baseline. We compare all these methods using classification accuracy on the held out test set. We repeat all the experiments 10 times.

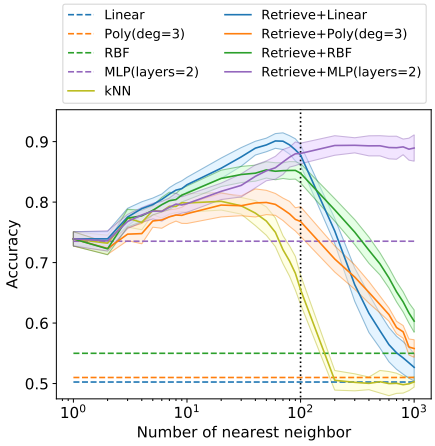



*Figure 4.* Performance of ERM and local ERM for various models on synthetic data.

**Observations.** In Figure 4, we observe the tradeoff of varying the size of the retrieved set (as dictated by the neighborhood radius) on the performance of the proposed algorithms. We see that when the number of retrieved samples is small the local methods have lower accuracy, this is due to large generalization error. When the size of the retrieved sample space is high, the local methods fail to minimize the loss effectively due to the lack of model capacity. We see that this effect being more pronounced for simpler function classes such as linear classifier as compared to RBF or polynomial classifiers.

## F.2. CIFAR-10

**Task and data.** We consider the task of binary classification on a *real image data* for object detection. In particular, we consider a subset of CIFAR-10 dataset where we only restrict to images from "Cat" and "Dog" classes. We randomly partition the data into a train set of $n = 10000$ points and remaining 2000 points for test. We do a 10-fold cross-validation.

**Methods.** We consider a subset of method from Appendix. F.1. In particular, we only consider a simple linear classifier and a multi-layer perceptron (MLP) of two layers. For retrieval-based models, we consider each of the above methods as the local model to fit on retrieved data points via local ERM framework (Sec. 3). The retrieval is done using L2 distance in the input space directly (no features is extracted). Additionally, we also report simple kNN baseline. We compare all these methods using classification accuracy on the held out test set. We repeat all the experiments 10 times.

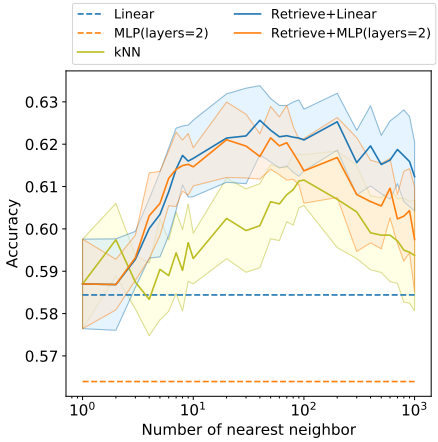



*Figure 5.* Performance of ERM and local ERM for various models on (binary) CIFAR-10.

**Observations.** Similar to Figure 4, Figure 5 exhibits a tradeoff, where varying the size of the retrieved set (as dictated by the neighborhood radius) impacts the performance of the proposed algorithms. We see when the number of retrieved samples is small the local methods have lower accuracy, this is due to large generalization error; and when the number of retrieved samples is large, simple local function class incurs a large approximation error.

## F.3. ImageNet

**Task and data.** We consider the task of 1000-way image classification on ImageNet ILSVRC-12 dataset. We use the standard train-test set split, where we have of $n = 1281167$ points for training and 50000 points for test. Given large computational cost, we could only run each experiment once.

**Methods.** We compare proposed Local ERM (Sec. 3) to state-of-the-art (SoTA) single model published for this task, which is from the most recent CVPR 2022 (Zhai et al., 2022). For the local parametric model we use a small MobileNetV3 architecture (Howard et al., 2019) with 4.01M parameters and 156 MFLOPs compute cost. Contrast this to SoTA model ViT-G/14 with 1.84B parameters and 938 GFLOPs compute cost. Following standard practice in literature, we use unsupervised learned features from ALIGN (Jia et al., 2021) to do image retrieval using L2 distance. For solving the local ERM, we fine-tune a MobileNetV3 model, which has been pretrained on ImageNet, on the retrieved set using Adam optimizer with a linear decay schedule. Additionally, we also report simple kNN baseline. We compare all these methods using classification accuracy on the held out test set.

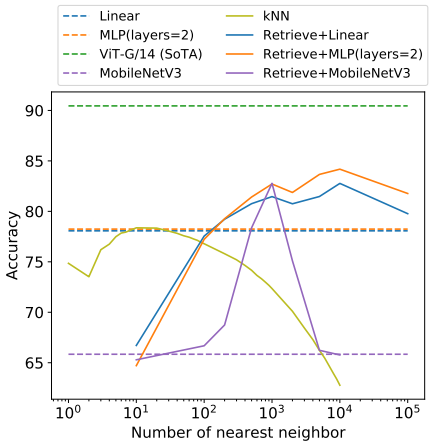



*Figure 6.* Performance of ERM and local ERM for various models on on ImageNet.

**Observations** In Figure 6, we see that local ERM with a small MobileNet-V3 model is able to achieve the top-1 accuracy of 82.78 whereas a regularly trained MobileNet-V3 model achieves the top-1 accuracy of only 65.80. Also the result is very competitive with SoTA of 90.45 with a *much larger model*. Thus, the result suggest that the simple local ERM framework (analyzed in our work) is able to demonstrate the utility of retrieval-based models. In particular, it allows a realistic small sized model to attain very competitive numbers on the popular ImageNet benchmark. Furthermore, as pointed at end of Sec. 3.4, using global representation from ALIGN embeddings help simplest linear model to outperform MobileNet-V3 working directly on image input, thereby showcasing the benefits of endowing local ERM with global representation.