Beyond Real Data: Synthetic Data Through The Lens Of Regularization

Amitis Shidani[1,2] **Tyler Farghly**[2]

Yang Sun[3] **Habib Ganjgahi**[3,†] **George Deligiannidis**[2,†]

[1]Apple [2]University of Oxford [3]Big Data Institute
[†] Joint senior supervision.
Correspondence to: amitis_shidani@apple.com

Abstract

Synthetic data can improve generalization when real data is scarce, but excessive reliance may introduce distributional mismatches that degrade performance. In this paper, we present a learning-theoretic framework to quantify the trade-off between synthetic and real data. Our approach leverages algorithmic stability to derive generalization error bounds, characterizing the optimal synthetic-to-real data ratio that minimizes expected test error as a function of the Wasserstein distance between the real and synthetic distributions. We motivate our framework in the setting of kernel ridge regression with mixed data, offering a detailed analysis that may be of independent interest. Our theory predicts the existence of an optimal ratio, leading to a U-shaped behavior of test error with respect to the proportion of synthetic data. Empirically, we validate this prediction on CIFAR-10 and a clinical brain MRI dataset. Our theory extends to the important scenario of domain adaptation, showing that carefully blending synthetic target data with limited source data can mitigate domain shift and enhance generalization. We conclude with practical guidance for applying our results to both in-domain and out-of-domain scenarios.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

The success of modern Machine Learning (ML) and Artificial Intelligence (AI) heavily depends on the availability of large-scale training datasets (Sun et al., 2017; Radford et al., 2021). However, in many critical domains such as healthcare, data collection is often prohibitively expensive, time-consuming, or constrained by privacy concerns (Esteva et al., 2019; Kaissis et al., 2021). Similar challenges arise in scientific domains where obtaining labeled data requires high-fidelity physical simulations or specialized experimental setups. For instance, generating data in molecular dynamics (Hollingsworth and Dror, 2018; Hansson et al., 2002) often demands significant computational resources, or structural biology techniques like cryo-electron microscopy (Murata and Wolf, 2018; Milne et al., 2013) involve costly and complex instrumentation. In these scenarios, ML models are trained on small datasets and as a result frequently suffer from poor generalization, limiting their practical applicability (Recht et al., 2019; Maleki et al., 2022; Schmidt et al., 2018; Brigato and Iocchi, 2021).

To address this challenge, several strategies have been proposed, including data augmentation (Shorten and Khoshgoftaar, 2019; Cubuk et al., 2020) and the use of synthetic data (Frid-Adar et al., 2018; Karras et al., 2020; Lu et al., 2023; Jordon et al., 2022). Although these methods can improve model accuracy, their success depends critically on how well the synthetic data approximates the real data distribution (Bowles et al., 2018). With the emergence of powerful generative models such as diffusion models (Ho et al., 2020; Song et al., 2021; Lipman et al., 2022; De Bortoli et al., 2021), there is renewed interest in using synthetic data to supplement limited real data (Trabucco et al., 2023;

Voetman et al., 2023; Alemohammad et al., 2024b). Empirical evidence suggests that, when properly generated, synthetic data can substantially boost the downstream model performance in low-data regimes (Azizi et al., 2023; Feng et al., 2024).

However, the integration of synthetic data introduces a critical trade-off as synthetic data may deviate from the true data distribution. If the synthetic dataset grows disproportionately large, the training algorithm may overlook the real data, introducing bias (Alemohammad et al., 2024a; Briesch et al., 2023; Betzalel et al., 2022; Dohmatob et al., 2025; Bertrand et al., 2024). See Appendix J for an extensive literature review on this topic. This issue motivates a central question:

"What is the optimal balance between real and synthetic data to minimize generalization error?"

In this work, we address this question from a learning-theoretic perspective, establishing that an optimal ratio of synthetic to real data exists for maximizing generalization performance. In Appendix D, we show that the traditional formalization of the problem fails to capture this trade-off, yielding a loose bound whose optimum lies at either excluding synthetic data entirely or using it without limit. Since this behaviour does not align with empirical observations (Section 4), we propose a modified formalization that more accurately reflects the practical balance. We first motivate our analysis through a simple yet insightful case study in kernel ridge regression (Singh and Vijaykumar, 2023), which may be of independent interest. We then extend our theoretical framework to more general settings, deriving generalization bounds via stability analysis (Bousquet and Elisseeff, 2002; Shalev-Shwartz and Ben-David, 2014; Hardt et al., 2016). Our theoretical insights are empirically validated on two distinct datasets: CIFAR-10 (a standard benchmark) (Krizhevsky et al., 2009) and a real-world brain imaging dataset for Multiple Sclerosis (MS) (Carass et al., 2017).

Furthermore, we extend our framework to domain adaptation settings (Ben-David et al., 2010; Ganin et al., 2016; Wilson and Cook, 2020), where synthetic data from a target domain is used to enhance limited real data from a source domain. This broadens the scope of our approach, highlighting its relevance for data-scarce scenarios in diverse ML applications like healthcare (Zhuang et al., 2021).

Contribution Our main contributions are as follows:

- We provide a learning-theoretic analysis demonstrating the existence of an optimal ratio between synthetic and real data that minimizes generalization error. Our approach is grounded in stability-based generalization bounds and is first illustrated through a tractable kernel ridge regression model. See Sections 2 and 3.
- We empirically validate our theoretical predictions using both benchmark (CIFAR-10, Appendix I.2) and real-world (brain MRI for Multiple Sclerosis, Section 4) datasets, confirming that an appropriate balance of synthetic data improves performance in low-data regimes.
- We extend our framework to domain adaptation, showing how synthetic data from a target domain can be effectively combined with limited real data from a source domain, thereby broadening the applicability of our results (Section 5). We also provide practical guidance for applying our theory to both in-domain and out-of-domain generalization tasks (Section 6).

2 MOTIVATION: SYNTHETIC DATA IN KERNEL REGRESSION

We study the effect of incorporating synthetic data into kernel regression (Singh and Vijaykumar, 2023; Allerbo, 2023; Wang and Jing, 2022; Smale and Zhou, 2005) as a simple yet illustrative setting to gain insight into the key factors influencing the generalization bound, i.e., when synthetic data improves or degrades generalization.

We consider kernel regression, where a function is learned by minimizing a regularized empirical risk over a Reproducing Kernel Hilbert Space (RKHS) denoted by $\mathcal{H}_K$. Given training data $\{(\mathrm{x}_n, \mathrm{y}_n)\}_{n=1}^N$ with $\mathrm{x}_n \in \mathcal{X} \sim p_{\mathbf{x}}$ and $\mathrm{y}_n \in \mathbb{R}$, the objective is to find a function $f \in \mathcal{H}_K$ that best fits the data while controlling complexity through a regularization term. We assume that $\mathrm{y}_n = f_\star(\mathrm{x}_n) + \varepsilon_n$, where ε_n are Independent and Identically Distributed (i.i.d.) samples from a zero-mean Gaussian distribution with variance σ^2. The Empirical Risk Minimization (ERM) normally takes the following form:

$$f_N = \arg\min_{f \in \mathcal{H}_k} \sum_{n=1}^{N} \left(\mathrm{y}_n - f(\mathrm{x}_n)\right)^2 + \sum_{m=1}^{M} \left(\tilde{\mathrm{y}}_m - f(\tilde{\mathrm{x}}_m)\right)^2 .$$

As shown in Appendix D, classical generalization arguments yield a loose bound on the test error of this formulation. The looseness arises mainly from the sample noise of synthetic data, which complicates the analysis. To address this, we approximate it with the following ansatz, where, unlike the standard setup, we regularize towards a synthetic data generator $g \in \mathcal{H}_K$, effectively corresponding to the case of having an infinite number of synthetic samples:

$$f_N = \arg\min_{f \in \mathcal{H}_k} \frac{1-\tilde{\lambda}}{N} \sum_{n=1}^{N} \left(\mathrm{y}_n - f(\mathrm{x}_n)\right)^2 + \tilde{\lambda}\|f - g\|^2 \quad (1)$$

where $\tilde{\lambda} > 0$ is the regularization strength. This perspective allows us to derive tighter bounds that align with empirical behavior. See Appendices D.2 and E for a detailed analysis of the asymptotics and the connection to the finite-sample formulation.

By the *Representer Theorem* (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001), the learned function takes the form $f_N(\mathrm{x}) = \sum_{n=1}^{N} \alpha_n K(\mathrm{x}, \mathrm{x}_n)$, where K is a positive definite kernel function, and α_n are coefficients obtained from a regularized least squares problem. Let $\tilde{\mathcal{H}} = \mathrm{span}\{K(\cdot, \mathrm{x}_1), \ldots, K(\cdot, \mathrm{x}_N)\}$, and $\mathcal{H}_K = \tilde{\mathcal{H}} \oplus \tilde{\mathcal{H}}^{\perp}$, where $\tilde{\mathcal{H}}^{\perp}$ is the orthogonal complement in $\tilde{\mathcal{H}}$. By the Representer Theorem, the synthetic data generator $g \in \mathcal{H}_K$ can be written as $g(\mathrm{x}) = \sum_{n=1}^{N} \beta_n K(\mathrm{x}, \mathrm{x}_n) + g_{\perp}(\mathrm{x})$, where $g_{\perp} \in \tilde{\mathcal{H}}^{\perp}$. Setting $g = 0$ recovers the standard kernel ridge regression. We establish the following lemma (proof in Appendix F.1), which characterizes the solution to Equation 1. Let $\lambda := \tilde{\lambda}/(1-\tilde{\lambda})$.

Lemma 2.1. *Let $K_N \in \mathbb{R}^{N \times N}$ be the empirical kernel matrix with entries $(K_N)_{ij} = K(\mathrm{x}_i, \mathrm{x}_j)$. Define the integral operator $T_K : L^2(p_x) \to L^2(p_x)$ by $(T_K f)(\mathrm{x}) = \int K(\mathrm{x}, x') f(x')\, dp_x(x') = \mathbb{E}_{\mathrm{x}'}[K(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')]$. Let $\lambda_N = N\lambda$. Then the solution to Equation 1 has the closed-form representation:*

$$\boldsymbol{\alpha} = (K_N + \lambda_N I)^{-1} (K_N \boldsymbol{\alpha}_\star + \lambda_N \boldsymbol{\beta} + \boldsymbol{\varepsilon}) ,$$

where $\boldsymbol{\alpha}_\star$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are the coefficients of $f_\star$, g, and the noise vector in the training basis.

We now recall the Mercer decomposition (Mercer, 1909) of the kernel. The operator T_K defined in Lemma 2.1 is compact, self-adjoint, and positive semi-definite, and thus admits a spectral decomposition. That is, there exist eigenfunctions $\{\phi_j\}_{j=1}^{\infty}$ forming an orthonormal basis of $L^2(p_x)$ and corresponding non-negative eigenvalues $\mu_1 \geq \mu_2 \geq \cdots \to 0$ such that $T_K \phi_j = \mu_j \phi_j$. The eigenfunctions ϕ_j can be interpreted as the natural coordinates of the function space with respect to the kernel, and the eigenvalues μ_j encode their relative importance. In this basis, we can write

$$f_\star = \sum_{j=1}^{\infty} \theta_j \phi_j , \qquad g = \sum_{j=1}^{\infty} \omega_j \phi_j , \tag{2}$$

where $\theta_j = \langle f_\star, \phi_j \rangle$, and $\omega_j = \langle g, \phi_j \rangle$.

Assumption 2.1 (Polynomial eigendecay and smoothness)**.** *We assume the kernel K exhibits $2r$-polynomial eigendecay for some $r \geq \frac{1}{2}$. Given the expansions in Equation 2, we assume for some $s, s' > 0$:*

$$\text{(a) } \theta_j^2 \asymp \mu_j^s \asymp j^{-2rs}, \text{ (b) } \omega_j^2 \asymp \mu_j^{s'} \asymp j^{-2rs'} .$$

Assumption 2.1 quantifies how well $f_\star$ and g align with the eigenfunctions of T_K. It ensures that $\sum_j \theta_j^2/\mu_j^s < \infty$ and $\sum_j \omega_j^2/\mu_j^{s'} < \infty$, i.e. $f_\star$ and g decay sufficiently fast in the eigenbasis; larger rate corresponds to greater smoothness. Such assumptions are standard in kernel regression analysis; see, e.g., Cheng et al. (2024); Bartlett et al. (2019); Cui et al. (2021); Barzilai and Shamir (2024).

Definition 2.1 (Bias-Variance Decomposition)**.** *Define the test error $\mathcal{R}_N(\lambda; g)$ to be the population mean squared error between the regressor and the true label averaged over noise:*

$$\mathcal{R}_N(\lambda; g) = \mathbb{E}_{\mathbf{x},\varepsilon}\left[(f_\star(\mathbf{x}) - f_N(\mathbf{x}))^2\right] .$$

We decompose the test error into a bias $\mathcal{B}$ and variance $\mathcal{V}$, with $\mathcal{R}_N(\lambda; g) = \mathcal{B}^2 + \mathcal{V}$, such that:

$$\mathcal{B}^2 = \mathbb{E}_{\mathbf{x}}\left[f_\star(\mathbf{x}) - \mathbb{E}_\varepsilon\left[f_N(\mathbf{x})\right]\right]^2 ,$$
$$\mathcal{V} = \mathbb{E}_{\mathbf{x},\varepsilon}\left[(f_N(\mathbf{x}) - \mathbb{E}_\varepsilon\left[f_N(\mathbf{x})\right])^2\right] .$$

We now present a bias–variance decomposition of Equation 1, along with a corollary characterizing the optimal number of synthetic samples. Proofs are in Appendices F.2 and F.3.

Theorem 2.2 (Generalization Error Bound)**.** *Under Assumption 2.1, for the kernel regression problem defined in Equation 1 and any fixed regularization parameter $\lambda > 0$, the test error admits the bound:*

$$\mathcal{R}_N(\lambda; g) = \mathcal{O}\left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\lambda^2} + \lambda^{2-\frac{1}{4r}} \mathcal{D}(f_\star, g)\right) ,$$

where $\lambda = \tilde{\lambda}/(1-\tilde{\lambda})$, and $\mathcal{D}(f_\star, g)^2 = \sum_{j=1}^{\infty} \frac{1}{\mu_j^2}(\theta_j - \omega_j)^2$ denotes the discrepancy between the target function $f_\star$ and the synthetic generator g.

Corollary 2.2.1 (Optimal Regularization and Synthetic Sample Size)**.** *Under the assumptions of Theorem 2.2, the optimal regularization parameter that minimizes the test error is given by*

$$\lambda^\star \asymp \left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} .$$

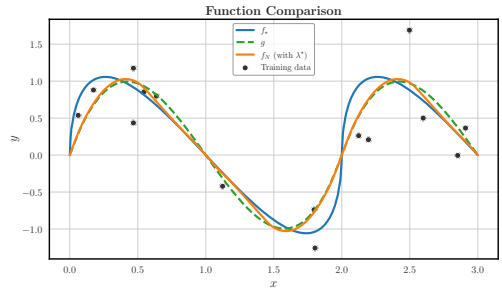
Setting $\lambda = \frac{M}{N}$, i.e., $\tilde{\lambda} = \frac{M}{N+M}$ (see Appendix D.2), the optimal number of synthetic samples satisfies:

$$M^\star \asymp \left(1 + \frac{\sigma^2}{\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} N^{\frac{12r+1}{16r+1}} .$$

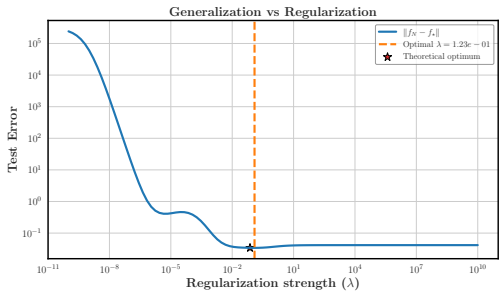
We empirically validate our theory in Figure 1, observing a U-curve as predicted by Theorem 2.2, with error minimized near the theoretical λ^*. See Appendix I.1 for details.

3 GENERALIZATION ERROR

We begin by introducing the notation and formal setting used throughout the remainder of the paper. We consider learning on a separable complete metric space



(a) Function comparisons.



(b) Generalization vs regularization.

Figure 1: *(a)* Comparison of the true function $f_\star$ (blue), the synthetic generator g (green), and the learned estimator f_N (orange), obtained via Lemma 2.1, with parameters $r = 2.0$, $s = 0.8$, and $s' = 1.5$. *(b)* Prediction error $|f_N - f_\star|_{L_2}$ as a function of the regularization strength λ. The U-shaped curve attains its minimum at $\lambda^\star$ (orange dashed line), which closely matches the theoretical optimum (star marker).

$(\mathcal{X}, d_{\mathcal{X}})$. We define the sample space $\mathcal{S}_N = \mathcal{X}^N$ and the random training dataset of N i.i.d. samples from p_{x} over $\mathcal{X}$ is denoted by $\mathbf{S} = \{\mathrm{x}_1, \ldots, \mathrm{x}_N\} \in \mathcal{S}_N$, with joint law $p_{\mathbf{S}}$. Consider some measurable hypothesis space $\mathcal{H}$ and a loss function $\ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}$ that quantifies the performance of a hypothesis, and we assume $\ell(h, \cdot) \in L^1(p_{\mathrm{x}})$ for each $h \in \mathcal{H}$. We define the empirical and population risks as,

$$\mathcal{L}_{\mathbf{S}}(h) = \frac{1}{N}\sum_{i=1}^{N} \ell(h, \mathrm{x}_i), \quad \mathcal{L}_{\mathcal{X}}(h) = r(h) = \mathbb{E}_{p_{\mathrm{x}}}[\ell(h, \mathrm{x})].$$

For $r \in [1, \infty)$, the Wasserstein r-distance between two probability measures p and q on $\mathcal{X}$ with finite r-moments is defined as $\mathcal{W}_r(p, q) = \inf_{\gamma \in \Gamma(p,q)} (\mathbb{E}_{(x,y)\sim\gamma}[d(x,y)^r])^{1/r}$, where $\Gamma(p, q)$ is the set of all couplings of p and q. See Appendix C for more detailed notation.

3.1 Synthetic Data As A Regularizer

Following the motivation in Section 2, we consider training on a mixture of real and synthetic data, where the synthetic data acts as a form of regularization. We focus on the following mixed loss:

$$\mathcal{R}_\lambda(h, \mathbf{S}) = (1 - \lambda)\mathcal{L}_{\mathbf{S}}(h) + \lambda \mathbb{E}_{\mathbf{x}\sim p'_{\mathbf{x}}}[\ell(h, \mathbf{x})],$$

where $p'_{\mathbf{x}}$ denotes the distribution of synthetic data, which may differ from the real distribution $p_{\mathbf{x}}$. We are interested in upper-bounding the generalization error of the algorithm that minimizes the mixed-loss. Our approach leverages a strategy from the learning theory literature known as algorithmic stability.

Definition 3.1 (Uniform Stability)**.** *Let $\mathcal{A} : \mathcal{S} \mapsto \mathcal{H}$ denote an algorithm. Algorithm $\mathcal{A}$ is ε-uniformly stable if for all $\mathbf{S}, \mathbf{S}' \in \mathcal{X}^N$ such that $\mathbf{S}, \mathbf{S}'$ differ in at most one example, the corresponding outputs $\mathcal{A}(\mathbf{S})$ and $\mathcal{A}(\mathbf{S}')$ satisfy $\sup_{\mathbf{x}\in\mathcal{X}} |\ell(\mathcal{A}(\mathbf{S}); \mathbf{x}) - \ell(\mathcal{A}(\mathbf{S}'); \mathbf{x})| \le \varepsilon$.*

This notion of algorithmic stability captures sensitivity of an algorithm on individual changes in the dataset. Under this property, it has been shown that generalization gap bounds in both expectation and high probability can be obtained (Bousquet and Elisseeff, 2000, 2002). In our analysis we consider the general case where $\mathcal{H}$ consists of set of functions between $\mathcal{X}$ and some metric space $\mathcal{Y}$. We make the assumption that it is a compact subset $L^\infty(p_{\mathrm{x}})$.

Assumption 3.1. *The hypothesis class $\mathcal{H}$ is a set of measurable functions of the form $\mathcal{X} \to \mathcal{Y}$ that are L-Lipschitz and, for some $D > 0$ satisfies $\|h - h'\|_{L^\infty(p_{\mathrm{x}})} \le D$, for any $h, h' \in \mathcal{H}$.*

Standard generalization bounds (e.g., Russo and Zou (2020); Lopez and Jog (2018); Clerico et al. (2022)) rely on regularity conditions on the loss function ℓ. We now recall the regularity conditions adopted in this work. We recall that a differentiable function $\phi : \mathcal{Y} \to \mathbb{R}$ is m-strongly convex for some constant $m > 0$ if it satisfies $\phi(x) \ge \phi(y) + \langle \nabla\phi(y), y - x\rangle + \frac{m}{2}\|x - y\|^2$ and is M-smooth if it satisfies $\phi(x) \le \phi(y) + \langle \nabla\phi(y), y - x\rangle + \frac{M}{2}\|x - y\|^2$.

Assumption 3.2. *The loss function takes the form $\ell(h, \mathrm{x}) = c(h(\mathrm{x}), \mathrm{x})$ for a function $c : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}^+$ that is M-smooth and, for every $\mathrm{x} \in \mathcal{X}$, the function $c(\cdot, \mathrm{x})$ is m-strongly convex and satisfies $\inf_{y\in\mathcal{Y}} c(y, \mathrm{x}) = 0$.*

This is satisfied by many common learning objectives, including regression with mean squared error and classification with cross-entropy loss. Furthermore, the use of smoothness and strong-convexity is standard within algorithmic stability and generalization (e.g., Bousquet and Elisseeff (2002, 2000); Charles and Papailiopoulos (2018); Bousquet et al. (2019); Yang et al. (2023); Shalev-Shwartz et al. (2010); Feldman and Vondrák (2019); Attia and Koren (2022); Farghly and Rebeschini (2021)).

Assumption 3.3. *Suppose that $\sigma := \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\|\mathbf{x}\|^2]^{1/2} < \infty$ and $\sigma' := \mathbb{E}_{\mathbf{x}\sim p'_{\mathbf{x}}}[\|\mathbf{x}\|^2]^{1/2} < \infty$*

and suppose that there exists $d^*, C > 0$ *such that for all* $x \in \mathrm{supp}(p_{\mathbf{x}})$ *and all sufficiently small* $\delta > 0$,

$$p_{\mathrm{x}}(B_\delta(x)) \geq C\delta^{d^*}.$$

We now state a result showing that the mixed-loss algorithm is uniformly stable and provides a bound on the generalization gap.

Theorem 3.1 (Mixed-data Generalization Bound)**.** *Suppose that Assumptions 3.1–3.3 hold, that* $h_{\mathbf{S}} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}_\lambda(h, \mathbf{S})$ *and set* $\mathcal{R}^\star = \min_{h\in\mathcal{H}} r(h)$. *Then, there exists* $N_0, \mathcal{R}_0^\star, \mathcal{W}_0 > 0$ *such that whenever* $N \geq N_0, \mathcal{R}^\star \leq \mathcal{R}_0^\star$ *and* $\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}) \leq \mathcal{W}_0$, *the algorithm* $\mathcal{A}(\mathbf{S}) = h_{\mathbf{S}}$ *is uniformly stable with constant,*

$$\varepsilon \lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(h_{\mathbf{S}})] + L^{\frac{2d^*}{d^*+2}} \Delta^{\frac{2}{d^*+2}},$$

$$\Delta = \frac{M}{Cm^2\lambda}\mathcal{R}^\star + \frac{\sqrt{M}D}{Cm\lambda N} + \frac{M\xi}{Cm^2}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}),$$

where ξ *is as defined in* (23). *Furthermore, we have the generalization bound,*

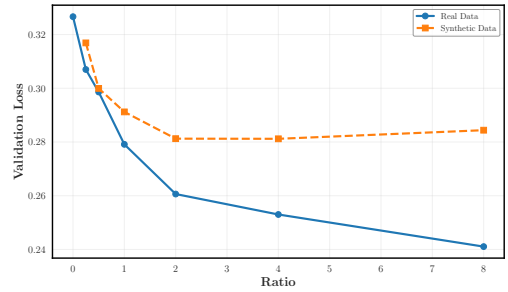
$$\mathbb{E}[r(h_{\mathbf{S}})] \lesssim \mathcal{R}^\star + \lambda\xi\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)L^{\frac{2d^*}{d^*+2}} \Delta^{\frac{2}{d^*+2}}.$$

The proof of Theorem 3.1, along with another result of stability for the mixed loss is discussed in Appendix G. The dimension d^* can be intuitively understood as the intrinsic dimension of the real data manifold. Notably, the generalization bound exhibits a U-shaped dependence on λ, similar to Theorem 2.2: for a fixed distributional discrepancy $\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}})$, there exists an optimal mixing parameter λ. This reflects a trade-off between algorithmic stability (which improves with more synthetic data) and distributional mismatch. In particular, when $\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}) = 0$, the optimal λ is 1, suggesting that it is beneficial to generate as much synthetic data as possible. We refer to the ratio $\frac{\lambda}{1-\lambda}$ as the *synthetic-to-real* ratio, which approximates $\frac{M}{N}$ in the finite-sample setting.

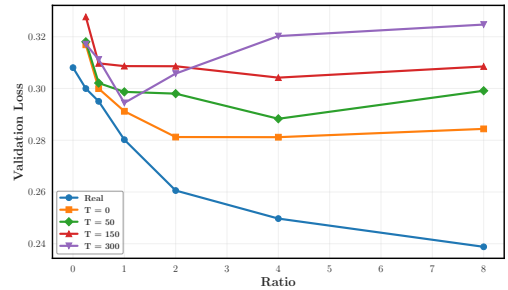
4 EXPERIMENTS: REAL-WORLD MEDICAL IMAGES

Multiple sclerosis (MS) is a chronic neurological disease affecting millions worldwide (Tullman, 2013). T2-hyperintense lesions in MRI reflect neuroinflammatory damage and serve as key biomarkers for diagnosis, monitoring, and prognosis (McGinley et al., 2021). Accurate segmentation of MS lesions in MRI remains a challenging problem. ML methods must contend with substantial variability in image characteristics, lesion appearance, and domain shift, arising from differences in scanners, acquisition protocols, and imaging parameters between training and test sets (Zeng et al., 2020). Furthermore, publicly available datasets for lesion segmentation remain limited in size and diversity, as acquiring labeled, heterogeneous MRI data is both costly and time-consuming.

Our work is motivated by the need to improve lesion segmentation performance under limited and heterogeneous training data and distributional shift. Specifically, we consider the setting where a synthetic data generator is available to address data scarcity and mitigate domain shift between the source (training) and target (test) domains, while potentially introducing additional distributional discrepancies. Following our theoretical result in Section 3, we study the in-domain setup in this section and refer to Section 5 for out-of-domain scenario.



(a) The effect of synthetic data.



(b) The effect of distributional distance.

Figure 2: *(a)* Validation loss decreases consistently as more real data is added (blue line), while increasing synthetic data (orange dashed line) produces a U-shaped curve, indicating an optimal mixing ratio λ, as predicted by Theorem 3.1. *(b)* Effect of distributional distance: varying the diffusion model timestep $T \in \{0, 50, 150, 300\}$ controls the noise level of synthetic samples. The U-shaped trend persists across all T but becomes sharper with increased discrepancy between real and synthetic distributions.

We conduct our experiments on the *NO.MS* dataset (Dahlke et al., 2021), one of the largest and most comprehensive clinical trial datasets for MS. It comprises over 200,000 MRI scans from more than 11,000 patients. Ground-truth lesion annotations are generated using an automated tool and subsequently refined by expert radiologists. The data originates from two Contract Research Organizations (CROs),

NeuroRx and MIAC, introducing inherent domain variability. For the downstream segmentation task, we use a training set of 100 NeuroRx scans (∼4,500 slices) and a fixed validation set of 20 NeuroRx scans (∼1,000 slices). In addition, we train a conditional diffusion model on the rest of NeuroRx data as our synthetic data generator. To empirically validate the theoretical insights from Theorem 3.1, we design two experiments:

1. **Effect of synthetic data:** We augment the training set by varying the synthetic-to-real ratio from 0.25 to 8, and compare performance against the ground truth, where the real dataset is scaled up.

2. **Distributional distance:** While we do not have direct access to the distance between the true and synthetic distributions, we study the effect of this discrepancy by varying the sampling timestep of the diffusion model ($T = 50$, 150, 300) out of 600 total denoising steps. We expect that samples from noisier timesteps exhibit greater distributional distance from the real data.

More details on the segmentation model architecture and hyperparameters are provided in Appendix I.3. Figure 2 shows that an appropriately chosen synthetic-to-real data ratio improves performance on the downstream segmentation task. Figure 2a compares validation loss when increasing the amount of synthetic data versus scaling up real data. As expected, adding more real data consistently improves performance. In contrast, synthetic data exhibits a U-shaped effect: moderate amounts enhance generalization, while excessive amounts degrade it, indicating the existence of an optimal interpolation parameter λ.

Figure 2b further examines how this behavior depends on the distributional distance between real and synthetic data. By varying the diffusion timestep T which controls the noise level in generated samples, we observe that the U-shape persists but becomes sharper as the synthetic data diverges further from the real distribution. These findings support our results that the generalization gap is influenced by both the mixing ratio and the distributional discrepancy between data sources. The relationship between the optimal synthetic-to-real ratio and distributional distance is further illustrated in Figure 8 in Appendix I.3.

5 DOMAIN ADAPTATION

In this section, we study the learning problem under *domain shift* (Zhang et al., 2019; Stacke et al., 2019; Redko et al., 2020; Shui et al., 2022): the real training data consist of samples from a source domain $\mathcal{X}$ with distribution $p_{\mathbf{x}}$, while the goal is to evaluate the learned model on a distinct target domain $\mathcal{X}^\star$ with distribution $p^\star_{\mathbf{x}}$, from which no real data are available. To address this distribution mismatch, we assume access to synthetic data generated on the target domain $\mathcal{X}^\star$, though drawn from a potentially imperfect distribution $p'_{\mathbf{x}} \neq p^\star_{\mathbf{x}}$. As in previous sections, this synthetic data is used to regularize the ERM objective, aiming to improve generalization to the target domain in the absence of real samples from $p^\star_{\mathbf{x}}$.

We first analyze this setting within the kernel framework (Section 2). Specifically, we consider a dataset of N real training pairs $\mathbf{y}_n = \tilde{f}(\mathbf{x}_n) + \varepsilon_n$, and a synthetic data generator g as defined earlier. The test error is measured with respect to a ground truth function $f_\star$. The main difference from Section 2 is that the training function $\tilde{f}$ differs from $f_\star$, capturing the domain shift. Although the empirical estimator remains unchanged (Equation 1), the generalization behavior is affected by the discrepancy between the training and target domains. Our result shows that stronger regularization can improve performance when the synthetic data more accurately approximates the target domain than the source data, providing a principled guideline for tuning λ under domain shift. See Appendix H.1 for the proof.

Theorem 5.1 (Generalization under Domain Shift)**.** *Under Assumption 2.1, for the kernel regression problem defined in Equation 1, distributional discrepancy $\mathcal{D}(\cdot,\cdot)$ as in Theorem 2.2, and any fixed regularization parameter $\lambda > 0$, the test error under domain shift satisfies the bound:*

$$\mathcal{R}_N(\lambda; g) \leq \frac{\mathcal{D}(f_\star, \tilde{f})}{\lambda^2} + \lambda^2 \mathcal{D}(f_\star, g) + \frac{\sigma^2 + \mathcal{D}(f_\star, \tilde{f}) + \mathcal{D}(f_\star, g)}{N\lambda^2}.$$

We now extend this result to the setup in Section 3, where test error is measured with respect to $p^\star_{\mathbf{x}}$. The resulting generalization gap is stated below; see Appendix H.2 for the proof.

Theorem 5.2 (Mixed-data Generalization under Domain Shift)**.** *Suppose that Assumptions 3.1–3.3 hold and that $h_{\mathbf{S}} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}_\lambda(h, \mathbf{S})$. Then, whenever $N \geq N_0, \mathcal{R}^\star \leq \mathcal{R}^\star_0$ and $\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}) \leq \mathcal{W}_0$, the generalization gap under the domain shift satisfies*

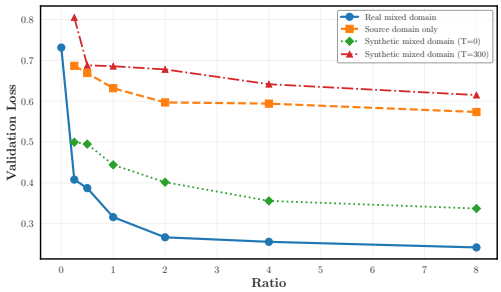
$$\mathbb{E}[r(h_{\mathbf{S}})] \lesssim r^\star + \lambda\xi\mathcal{W}_2(p^\star_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)\xi\mathcal{W}_2(p^\star_{\mathbf{x}}, p_{\mathbf{x}}) + (1-\lambda)L^{\frac{2d^*}{d^*+2}}\Delta^{\frac{2}{d^*+2}},$$

where $r^\star = \min_{h\in\mathcal{H}} r^\star(h)$ is minimum population risk minimizer of the target domain and $\Delta, \xi, N_0, \mathcal{R}^\star_0, \mathcal{W}_0$ are as in Theorem 3.1.

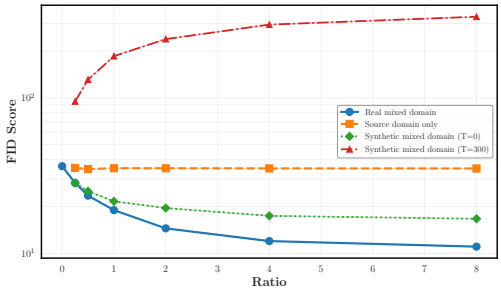
As expected, compared to Theorems 2.2 and 3.1, these bounds include an additional term that captures the mismatch between the source and target distributions. Consequently, the optimal choice of λ depends on the relative magnitudes of $\mathcal{D}(f_\star, \tilde{f})$ and $\mathcal{D}(f_\star, g)$ or similarly $\mathcal{W}_2(p^\star_{\mathbf{x}}, p_{\mathbf{x}})$ and $\mathcal{W}_2(p^\star_{\mathbf{x}}, p'_{\mathbf{x}})$. Intuitively, when

the synthetic data generator more closely approximates the target domain, it is beneficial to choose a larger regularization parameter λ. In the special case where $f_\star = \tilde{f}$ or $\mathcal{W}_2\left(p^\star_{\mathbf{x}}, p_{\mathbf{x}}\right) = 0$, the bound reduces to the previous results, albeit with potentially larger constants for the kernel regression case, arising from the more general proof strategy.

Experimental setup We follow the experimental setup for medical brain MRI scans described in Section 4 to study the effect of domain shift. Since we have two data sources (MIAC and NeuroRx), we can naturally adapt the setup to introduce domain shift: we treat MIAC as the source domain and NeuroRx as the target domain. The synthetic data generator, a conditional diffusion model, is trained on NeuroRx, thus approximating the target distribution. As before, we vary the synthetic-to-real data ratio in the range 0.25 to 8, and compare the resulting performance. Results are shown in Figure 3a. We include two baselines in this experiment: (1) access to real data from the target domain for training the downstream segmentation task on NeuroRx (blue line), and (2) no access to either target or synthetic data, with only increased source domain data available (orange line). To examine the impact of distributional discrepancy between synthetic and target data, we adopt the same approach as in Section 4, sampling from the diffusion model at two timesteps, $T \in \{0, 300\}$. We expect $T = 0$ (green dashed line) to closely match the target distribution, while $T = 300$ (red dashed line) reflects a greater distributional distance. As observed, synthetic data can significantly improve performance when the distributional distance between the synthetic and target is small. However, when the synthetic generator induces a large distributional shift, using additional source data alone can be more effective, but if the source domain itself is far from the target, neither synthetic nor source data is likely to help. This observation aligns with our theoretical understanding of the trade-offs in generalization error, where the benefit of additional data depends critically on the distributional closeness to the target domain.



(a) The effect of synthetic data in domain shift.



(b) FID score vs. synthetic-to-real data ratio

Figure 3: *(a)* Effect of synthetic data from distributions close to (green dashed) or far from (red dashed) the target, compared to target (blue) and source (orange) baselines. Results show the trade-off between distributional shift and regularization predicted by Theorem 5.2. *(b)* FID as a proxy for distributional shift: $T = 0$ (green) aligns with the target, while noisy (red) and source (orange) data show higher FID and reduced utility.

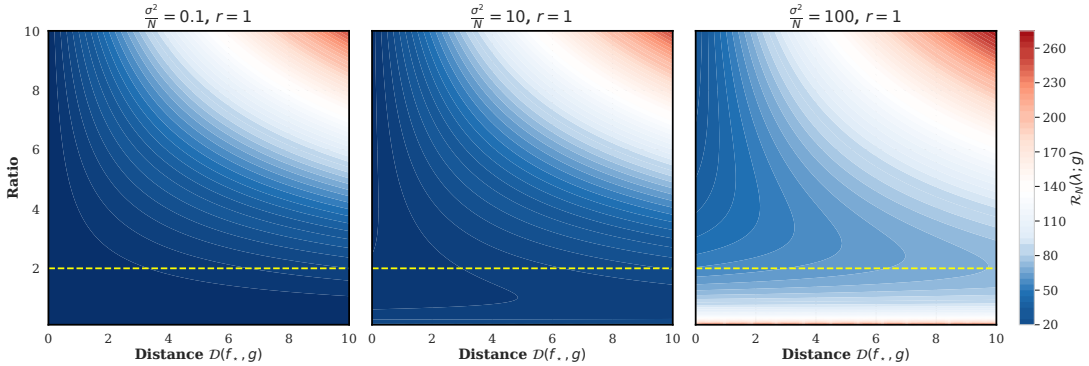
6 INSIGHTS FOR PRACTITIONERS

To apply our theoretical results, practitioners must estimate key quantities affecting the generalization bound, such as distributional distances, noise levels, and hypothesis class complexity. This section outlines practical ways to approximate these quantities and offers heuristics based on empirical evidence.

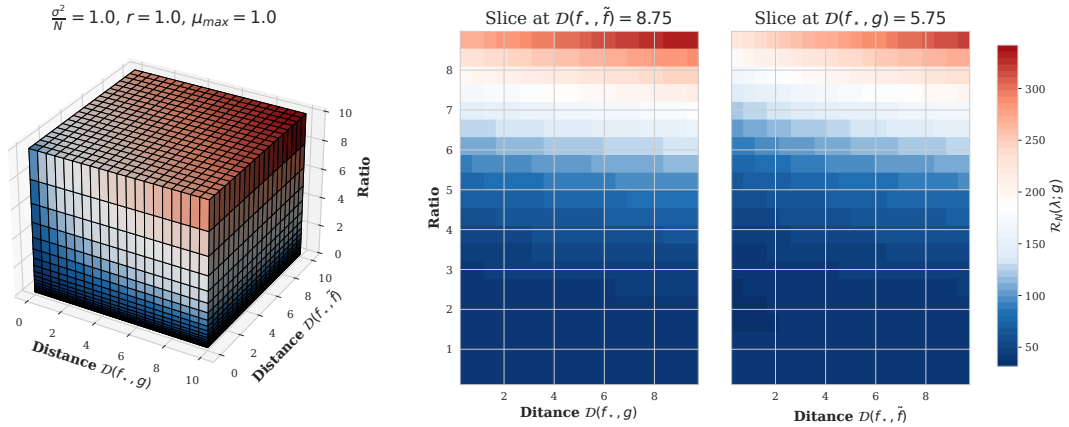
Exact distributional distances between real and synthetic data, or across domains, are rarely available. Applications therefore rely on proxies. In our experiments (Sections 4 and 5), we first used diffusion timesteps, and here adopt the widely used Fréchet Inception Distance (FID) metric. While FID is not a true distance, it approximates distributional alignment via the Wasserstein distance between Gaussians fitted to Inception embeddings of real and generated images. When synthetic data is well aligned with the target domain, FID provides a useful indicator for comparing generators or estimating alignment. As shown in Figure 3b, adding synthetic data from $T = 0$ (green) reduces FID similarly to adding real target-domain data (blue), mirroring trends in Figure 3a. In contrast, source-domain or noisy diffusion samples do not improve—and often worsen—FID. Depending on context, alternatives such as cross-validation loss or Kullback-Leibler Divergence (KLD) may also be viable.

We also propose a more accurate estimate based on Corollary 2.2.1. In image experiments, we exploit frequency-domain features: a 2D Fourier transform preserves information while decomposing energy across frequencies, analogous to the kernel eigenbasis and suitable for analyzing eigendecay.

- **Distributional distance.** Assuming both real and synthetic images are from related distributions



(a) Contour plot illustrating the behavior described in Theorem 2.2.



(b) Contour plot illustrating the behavior described in Theorem 5.1.

Figure 4: Effect of synthetic-to-real data ratio and distributional distance(s) on the error rate: *(a)* in-domain scenario across various signal-to-noise ratios for the real dataset; *(b)* out-of-domain scenario. In both cases, one should ideally choose λ such that it lies within the blue regions, which correspond to lower error rates.

(possibly with a shift), we compute the radially averaged power spectral density (RAPSD) for each distribution. We then estimate the distributional distance as the ℓ_2-distance between the average RAPSD vectors.

- **Decay exponent.** The eigendecay exponent r is estimated from the slope of a log-log plot of power spectrum (Q) vs. frequency (ω):

$$Q \asymp \omega^{-2r} \quad \Rightarrow \quad \log Q \approx -2r \log \omega + \text{const}.$$

We implement this on a real dataset, where the distance between real and synthetic data at $T = 0$ is 0.85, which we use for our experiments.

Another critical factor is data variance. This can be tricky, as many datasets are normalized to unit variance. In our setup, the variance in RGB setup yields $\sigma = 251.4$, giving $\lambda^\star = 1.958$, close to the observed ratio of 2. While normalization yields $\lambda^\star = 0.12$. This highlights the sensitivity of our result to the scale of variance and is one of the limitations of our method.

Figure 4 illustrates how the bound varies across scenarios, providing heuristics for mixing real and synthetic data. Lower bounds (cooler colors) indicate better generalization; higher ones (red) warn of overfitting. See Appendix I.4 for extended results.

In-domain. When heterogeneity is small to moderate and synthetic quality is high ($D(f_*, g)$ small), augmenting with up to twice as much synthetic as real data is effective (Figure 4a). Beyond this, gains diminish, with higher computational cost but little improvement. In biomedical cases (e.g., brain MRI) with small N and high σ^2, a 1:2 ratio still works well, though further increases only help with exceptionally accurate generators. Otherwise, collecting more real data is preferable.

Out-of-domain. Under domain shift (Figure 4b), the same 1:2 ratio is robust if the generator is good. Larger ratios harm performance, especially under severe shift ($\mathcal{D}(f_\star, \tilde{f}) \approx 8.75$). For moderate shifts ($\mathcal{D}(f_\star, g) \approx 5$), 1:2 remains a reliable choice. Overall, ratios between 1:1 and 1:2 seem to be effective. Modest augmentation helps even with shift, but excessive synthetic data

misaligned with the target distribution degrades performance. Careful tuning of the augmentation ratio is thus crucial in the out-of-domain case.

Finally, our theory assumes Lipschitz continuity, which can be estimated via gradient norms or enforced by clipping. This only scales the bound by a constant and does not affect the optimal ratio order.

7 CONCLUSION

Synthetic data is vital in domains where real data is scarce, costly, or sensitive, such as healthcare. We develop a principled framework that characterizes the trade-off between real and synthetic data and prove the existence of an optimal synthetic-to-real ratio that minimizes generalization error, first in kernel regression and then more generally via stability. Experiments on benchmarks and real-world datasets validate these predictions, revealing a non-monotonic relationship between performance and synthetic proportion and extending to domain adaptation with distribution shift.

We show that most traditional, model-agnostic techniques that rely on uniform bounds are often loose and unable to capture the phenomena observed in practice. Our modified ERM objective relaxes these assumptions by ignoring the variance of synthetic data, and yields tighter guarantees. However, a deeper understanding is still needed; for instance, PAC-Bayes bounds that treat synthetic data as a prior may provide a promising next step.

Acknowledgments

We thank Russ Webb and Dieter Haering for their useful feedback.

Bibliography

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. (2024a). Self-consuming generative models go MAD. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Alemohammad, S., Humayun, A. I., Agarwal, S., Collomosse, J., and Baraniuk, R. (2024b). Self-improving diffusion models with synthetic data. *arXiv preprint arXiv:2408.16333*.

Allerbo, O. (2023). Solving kernel ridge regression with gradient descent for a non-constant kernel. *arXiv preprint arXiv:2311.01762*.

Anthony, M. and Bartlett, P. L. (2002). *Neural Network Learning - Theoretical Foundations*. Cambridge University Press.

Asadi, A., Abbe, E., and Verdú, S. (2018). Chaining mutual information and tightening generalization bounds. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7245–7254.

Attia, A. and Koren, T. (2022). Uniform stability for first-order empirical risk minimization. In Loh, P. and Raginsky, M., editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 3313–3332. PMLR.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., Chen, T., Tomasev, N., Mitrović, J., Strachan, P., et al. (2023). Synthetic data in healthcare. *arXiv preprint arXiv:2306.08037*.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *CoRR*, abs/1906.11300.

Barzilai, D. and Shamir, O. (2024). Generalization in kernel regression under realistic assumptions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Belkin, M., Ma, S., and Mandal, S. (2018). To understand deep learning we need to understand kernel learning. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 540–548. PMLR.

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

Bertrand, Q., Bose, A. J., Duplessis, A., Jiralerspong, M., and Gidel, G. (2024). On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Betzalel, E., Penso, C., Navon, A., and Fetaya, E. (2022). A study on the evaluation of generative models. *arXiv preprint arXiv:2206.10935*.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In Bousquet, O., von Luxburg, U., and Rätsch, G., editors,

Advanced Lectures on Machine Learning, ML Summer Schools 2003, Canberra, Australia, February 2-14, 2003, Tübingen, Germany, August 4-16, 2003, Revised Lectures, volume 3176 of *Lecture Notes in Computer Science*, pages 169–207. Springer.

Bousquet, O. and Elisseeff, A. (2000). Algorithmic stability and generalization performance. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, pages 196–202. MIT Press.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *J. Mach. Learn. Res.*, 2:499–526.

Bousquet, O., Klochkov, Y., and Zhivotovskiy, N. (2019). Sharper bounds for uniformly stable algorithms. *CoRR*, abs/1910.07833.

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D. A., Hernández, M. V., Wardlaw, J., and Rueckert, D. (2018). Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*.

Briesch, M., Sobania, D., and Rothlauf, F. (2023). Large language models suffer from their own output: An analysis of the self-consuming training loop. *CoRR*, abs/2311.16822.

Brigato, L. and Iocchi, L. (2021). A close look at deep learning with small data. In *2020 25th international conference on pattern recognition (ICPR)*, pages 2490–2497. IEEE.

Busbridge, D., Shidani, A., Weers, F., Ramapuram, J., Littwin, E., and Webb, R. (2025). Distillation scaling laws. In *Forty-second International Conference on Machine Learning*.

Carass, A., Roy, S., Jog, A., Cuzzocreo, J. L., Sweeney, E. M., Gherman, A., Button, J., Nguyen, J., Prados, F., Sudre, C. H., Cardoso, M. J., Cawley, N., Ciccarelli, O., Wheeler-Kingshott, C. A. M., Ourselin, S., Catanese, L., Deshpande, H., Maurel, P., Commowick, O., Barillot, C., and Tomas-Fernandez, X. (2017). Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 148:77–102.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9630–9640. IEEE.

Charles, Z. and Papailiopoulos, D. S. (2018). Stability and generalization of learning algorithms that converge to global optima. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 744–753. PMLR.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pages 1597–1607. PMLR.

Cheng, T. S., Lucchi, A., Kratsios, A., and Belius, D. (2024). A comprehensive analysis on the learning curve in kernel ridge regression. *CoRR*, abs/2410.17796.

Clerico, E., Shidani, A., Deligiannidis, G., and Doucet, A. (2022). Chained generalisation bounds. In Loh, P. and Raginsky, M., editors, *Conference on Learning Theory, 2-5 July 2022, London, UK*, volume 178 of *Proceedings of Machine Learning Research*, pages 4212–4257. PMLR.

Cortés, A., Rodríguez, C., Vélez, G., Barandiarán, J., and Nieto, M. (2024). Analysis of classifier training on synthetic data for cross-domain datasets. *CoRR*, abs/2410.22748.

Cubuk, E. D., Zoph, B., Shlens, J., and Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. (2021). Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *CoRR*, abs/2105.15004.

Dahlke, F., Arnold, D. L., Aarden, P., Ganjgahi, H., Häring, D. A., Čuklina, J., Nichols, T. E., Gardiner, S., Bermel, R., and Wiendl, H. (2021). Characterisation of MS phenotypes across the age span using a novel data set integrating 34 clinical trials (NO.MS cohort): Age is a key contributor to presentation. *Multiple Sclerosis Journal*, 27(13):2062–2076.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. (2021). Diffusion schrödinger bridge with applications to score-based generative modeling. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17695–17709. Curran Associates, Inc.

Dohmatob, E., Feng, Y., and Kempe, J. (2024). Model collapse demystified: The case of regression. *CoRR*, abs/2402.07712.

Dohmatob, E., Feng, Y., Subramonian, A., and Kempe, J. (2025). Strong model collapse. In *The Thirteenth International Conference on Learning Representations*.

Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., and Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29.

Farghly, T. and Rebeschini, P. (2021). Time-independent generalization bounds for sgld in non-convex settings. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 19836–19846. Curran Associates, Inc.

Farghly, T., Rebeschini, P., Deligiannidis, G., and Doucet, A. (2026). Implicit regularisation in diffusion models: An algorithm-dependent generalisation analysis. In *The Fourteenth International Conference on Learning Representations*.

Feldman, V. and Vondrák, J. (2019). High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. *CoRR*, abs/1902.10710.

Feng, Y., Dohmatob, E., Yang, P., Charton, F., and Kempe, J. (2024). Beyond model collapse: Scaling up with synthesized data requires verification. *arXiv preprint arXiv:2406.07515*.

Ferbach, D., Bertrand, Q., Bose, A. J., and Gidel, G. (2024). Self-consuming generative models with curated data provably optimize human preferences. *CoRR*, abs/2407.09499.

Firdoussi, A. E., Seddik, M. E. A., Hayou, S., ALAMI, R., Alzubaidi, A., and Hacid, H. (2025). Synthetic data pruning in high dimensions: A random matrix perspective. In *Will Synthetic Data Finally Solve the Data Access Problem?*

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *arXiv preprint arXiv:1803.01229*.

Gálvez, B. R., Bassi, G., Thobaben, R., and Skoglund, M. (2020). On random subset generalization error bounds and the stochastic gradient langevin dynamics algorithm. *CoRR*, abs/2010.10994.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Garg, A., Bhattacharya, S., and Sur, P. (2025). Preventing model collapse under overparametrization: Optimal mixing ratios for interpolation learning and ridge regression. *arXiv preprint arXiv:2509.22341*.

Gerace, F., Saglietti, L., Mannelli, S. S., Saxe, A. M., and Zdeborová, L. (2022). Probing transfer learning with a model of synthetic correlated datasets. *Mach. Learn. Sci. Technol.*, 3(1):15030.

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. (2024). Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *CoRR*, abs/2404.01413.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. (2012). A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

Guedj, B. (2019). A primer on pac-bayesian learning. *CoRR*, abs/1901.05353.

Haghifam, M., Negrea, J., Khisti, A., Roy, D. M., and Dziugaite, G. K. (2020). Sharpened generalization bounds based on conditional mutual information and an application to noisy, iterative algorithms. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hansson, T., Oostenbrink, C., and van Gunsteren, W. (2002). Molecular dynamics simulations. *Current opinion in structural biology*, 12(2):190–196.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In Balcan, M. F. and Weinberger, K. Q., editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1225–1234, New York, New York, USA. PMLR.

Hashimoto, T. (2021). Model performance scaling with multiple data sources. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4107–4116. PMLR.

He, H., Xu, S., and Cheng, G. (2025). Golden ratio weighting prevents model collapse. *arXiv preprint arXiv:2502.18049*.

Hinton, G. E., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.

Hollingsworth, S. A. and Dror, R. O. (2018). Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143.

Hou, S., Kassraie, P., Kratsios, A., Krause, A., and Rothfuss, J. (2023). Instance-dependent generalization bounds via optimal transport. *J. Mach. Learn. Res.*, 24:349:1–349:51.

Imbusch, B. T., Schwarz, M., and Behnke, S. (2022). Synthetic-to-real domain adaptation using contrastive unpaired translation. In *18th IEEE International Conference on Automation Science and Engineering, CASE 2022, Mexico City, Mexico, August 20-24, 2022*, pages 595–602. IEEE.

Jain, A., Montanari, A., and Sasoglu, E. (2024). Scaling laws for learning with real and surrogate data. *CoRR*, abs/2402.04376.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. (2022). Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*.

Kaissis, G., Ziller, A., Passerat-Palmbach, J., Ryffel, T., Usynin, D., Trask, A., Lima Jr, I., Mancuso, J., Jungmann, F., Steinborn, M.-M., et al. (2021). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 3(6):473–484.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Kim, Y., Soh, J. W., Park, G. Y., and Cho, N. I. (2020). Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3479–3489. Computer Vision Foundation / IEEE.

Kimeldorf, G. and Wahba, G. (1971). Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, Citeseer.

Lee, H., Zhang, Y., Kwon, H., and Bhattacharrya, S. S. (2024). Exploring the potential of synthetic data to replace real data. *CoRR*, abs/2408.14559.

Li, D. and Zhang, H. R. (2021). Improved regularization and robustness for fine-tuning in neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27249–27262.

Lin, C.-H., Kaushik, C., Dyer, E. L., and Muthukumar, V. (2022). The good, the bad, and the ugly sides of data augmentation: An implicit spectral regularization perspective. *arXiv preprint arXiv:2210.05021*.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. (2022). Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Lopez, A. T. and Jog, V. S. (2018). Generalization error bounds using wasserstein distances. In *IEEE Information Theory Workshop, ITW 2018, Guangzhou, China, November 25-29, 2018*, pages 1–5. IEEE.

Lu, Y., Shen, M., Wang, H., Wang, X., van Rechem, C., Fu, T., and Wei, W. (2023). Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*.

Maleki, F., Ovens, K., Gupta, R., Reinhold, C., Spatz, A., and Forghani, R. (2022). Generalizability of machine learning models: quantitative evaluation of three methodological pitfalls. *Radiology: Artificial Intelligence*, 5(1):e220028.

McAllester, D. A. (1999). Pac-bayesian model averaging. In Ben-David, S. and Long, P. M., editors, *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, COLT 1999, Santa Cruz, CA, USA, July 7-9, 1999*, pages 164–170. ACM.

McGinley, M. P., Goldschmidt, C. H., and Rae-Grant, A. D. (2021). Diagnosis and treatment of multiple sclerosis: a review. *Jama*, 325(8):765–779.

Menon, A. K., Rawat, A. S., Reddi, S. J., Kim, S., and Kumar, S. (2020). Why distillation helps: a statistical perspective. *CoRR*, abs/2005.10419.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Phil. Trans. Roy. Soc. Lond. A*, 209:415.

Milne, J. L., Borgnia, M. J., Bartesaghi, A., Tran, E. E., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., and Subramaniam, S. (2013). Cryo-electron microscopy–a primer for the non-microscopist. *The FEBS journal*, 280(1):28–45.

Mishra, S., Panda, R., Phoo, C. P., Chen, C. R., Karlinsky, L., Saenko, K., Saligrama, V., and Feris, R. S. (2022). Task2sim: Towards effective pre-training and transfer from synthetic data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 9184–9194. IEEE.

Misiakiewicz, T. and Saeed, B. (2024). A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and GCV estimator. *CoRR*, abs/2403.08938.

Mou, W., Zhou, Y., Gao, J., and Wang, L. (2018). Dropout training, data-dependent regularization, and generalization bounds. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3642–3650. PMLR.

Mullick, K., Jain, H., Gupta, S., and Kale, A. A. (2023). Domain adaptation of synthetic driving datasets for real-world autonomous driving. *CoRR*, abs/2302.04149.

Murata, K. and Wolf, M. (2018). Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1862(2):324–334. Biophysical Exploration of Dynamical Ordering of Biomolecular Systems.

Peng, X., Usman, B., Saito, K., Kaushik, N., Hoffman, J., and Saenko, K. (2018). Syn2real: A new benchmark for synthetic-to-real visual domain adaptation. *CoRR*, abs/1806.09755.

Polyanskiy, Y. and Wu, Y. (2015). Wasserstein continuity of entropy and outer bounds for interference channels. *CoRR*, abs/1504.04419.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Raginsky, M., Rakhlin, A., and Telgarsky, M. (2017). Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. *CoRR*, abs/1702.03849.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*.

Redko, I., Morvant, E., Habrard, A., Sebban, M., and Bennani, Y. (2020). A survey on domain adaptation theory: learning bounds and theoretical guarantees. *arXiv preprint arXiv:2004.11829*.

Russo, D. and Zou, J. (2020). How much does your data exploration overfit? controlling bias via information usage. *IEEE Trans. Inf. Theory*, 66(1):302–323.

Saberi, A. H., Najafi, A., Emrani, A., Behjati, A., Zolfimoselo, Y., Shadrooy, M., Motahari, A. S., and Khalaj, B. H. (2024a). Gradual domain adaptation via manifold-constrained distributionally robust optimization. *CoRR*, abs/2410.14061.

Saberi, S. A. H., Najafi, A., Heidari, A., Movasaghinia, M. H., Motahari, A. S., and Khalaj, B. H. (2024b). Out-of-domain unlabeled data improves generalization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Sariyildiz, M. B., Alahari, K., Larlus, D., and Kalantidis, Y. (2023). Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 8011–8021. IEEE.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. (2018). Adversarially robust generalization requires more data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In Helmbold, D. P. and Williamson, R. C., editors, *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, volume 2111 of *Lecture Notes in Computer Science*, pages 416–426. Springer.

Seib, V., Lange, B., and Wirtz, S. (2020). Mixing real and synthetic data to enhance neural network training - A review of current approaches. *CoRR*, abs/2007.08781.

Shakeri, S., dos Santos, C. N., Zhu, H., Ng, P., Nan, F., Wang, Z., Nallapati, R., and Xiang, B. (2020). End-to-end synthetic data generation for domain adaptation of question answering systems. *CoRR*, abs/2010.06028.

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

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. (2010). Learnability, stability and uniform convergence. *J. Mach. Learn. Res.*, 11:2635–2670.

Shen, M., Bu, Y., and Wornell, G. W. (2023). On balancing bias and variance in unsupervised multi-source-free domain adaptation. *arXiv preprint arXiv:2301.13228*.

Shidani, A., Hjelm, R. D., Ramapuram, J., Webb, R., Dhekane, E. G., and Busbridge, D. (2024). Poly-view contrastive learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2017). Learning from simulated and unsupervised images through adversarial training. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2242–2251. IEEE Computer Society.

Shui, C., Chen, Q., Wen, J., Zhou, F., Gagné, C., and Wang, B. (2022). A novel domain adaptation theory with jensen–shannon divergence. *Knowledge-Based Systems*, 257:109808.

Shukor, M., Bethune, L., Busbridge, D., Grangier, D., Fini, E., El-Nouby, A., and Ablin, P. (2025). Scaling laws for optimal data mixtures. *arXiv preprint arXiv:2507.09404*.

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. J., and Gal, Y. (2024). AI models collapse when trained on recursively generated data. *Nat.*, 631(8022):755–759.

Singh, R. and Vijaykumar, S. (2023). Kernel ridge regression inference. *arXiv preprint arXiv:2302.06578*.

Smale, S. and Zhou, D.-X. (2005). Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3):285–302. Computational Harmonic Analysis - Part 1.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Stacke, K., Eilertsen, G., Unger, J., and Lundström, C. (2019). A closer look at domain shift for deep learning in histopathology. *arXiv preprint arXiv:1909.11575*.

Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A. A., and Wilson, A. G. (2021). Does knowledge distillation really work? In Ranzato, M., Beygelzimer, A., Dauphin, Y. N., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 6906–6919.

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. *arXiv preprint arXiv:1707.02968*.

Thudi, A., Rovers, E., Ruan, Y., Thrush, T., and Maddison, C. J. (2025). Mixmin: Finding data mixtures via convex minimization. *arXiv preprint arXiv:2502.10510*.

Trabucco, B., Doherty, K., Gurinas, M., and Salakhutdinov, R. (2023). Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Tullman, M. J. (2013). Overview of the epidemiology, diagnosis, and disease progression associated with multiple sclerosis. *Am J Manag Care*, 19(2 Suppl):S15–20.

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory, Second Edition*. Statistics for Engineering and Information Science. Springer.

Voetman, R., Aghaei, M., and Dijkstra, K. (2023). The big data myth: Using diffusion models for dataset generation to train deep detection models. *arXiv preprint arXiv:2306.09762*.

Wang, W. and Jing, B.-Y. (2022). Gaussian process regression: Optimality, robustness, and relationship with kernel ridge regression. *Journal of Machine Learning Research*, 23(193):1–67.

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46.

Wu, Q., Li, J. Y.-M., and Mao, T. (2022). On generalization and regularization via wasserstein distributionally robust optimization. *arXiv preprint arXiv:2212.05716*.

Yang, M., Wei, X., Yang, T., and Ying, Y. (2023). Stability and generalization of stochastic compositional gradient descent algorithms. *CoRR*, abs/2307.03357.

Ye, J., Liu, P., Sun, T., Zhan, J., Zhou, Y., and Qiu, X. (2025). Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *The Thirteenth International Conference on Learning Representations*.

Zeng, C., Gu, L., Liu, Z., and Zhao, S. (2020). Review of deep learning approaches for the segmentation of multiple sclerosis lesions on brain mri. *Frontiers in Neuroinformatics*, 14:610967.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Zhang, S., Luo, Y., Wang, Q., Chi, H., Chen, X., Han, B., and Li, J. (2023). Mixture data for training cannot ensure out-of-distribution generalization. *arXiv preprint arXiv:2312.16243*.

Zhang, X., Fu, Y., Zang, A., Sigal, L., and Agam, G. (2015). Learning classifiers from synthetic data using a multichannel autoencoder. *CoRR*, abs/1503.03163.

Zhang, Y., Liu, T., Long, M., and Jordan, M. (2019). Bridging theory and algorithm for domain adaptation. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7404–7413. PMLR.

Zhezherau, A. and Yanockin, A. (2024). Hybrid training approaches for llms: Leveraging real and synthetic data to enhance model performance in domain-specific applications. *arXiv preprint arXiv:2410.09168*.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, see Sections 2 and 3, and Appendix C for the mathematical setup.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, the main results and analysis are in Theorems 2.2 and 5.1 for the kernel regression case and Theorems 3.1 and 5.2 for the general case.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. No, experiments primarily serve to illustrate our theoretical findings. We are happy to release the code upon request by the reviewers.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, see Assumption 2.1, Assumption 3.1 and Assumption 3.2.
   (b) Complete proofs of all theoretical results. Yes, proofs are available in Appendices F.1 and H.1 and Appendices G.3 and H.2.
   (c) Clear explanations of any assumptions. Yes, see Assumption 2.1, Assumption 3.1 and Assumption 3.2 for references.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). No, experiments primarily serve to illustrate our theoretical findings. We are happy to release the code upon request by the reviewers.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes, all experimental details are provided in Section 4 and Appendices I.1 and I.2.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, see Figures 9a to 9c for the error bar and averaged runs over three seeds.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes, the compute budget and infrastructure is explained in Appendix I.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes, the description of the data along with the related references is available in Section 4 and Appendix I.2).
   (b) The license information of the assets, if applicable. Yes. For NO.MS, the raw data (anonymized) and associated documents (e.g., protocol, reporting and analysis plan, clinical study report) can be requested via https://www.clinicalstudydatarequest.com by signing a Data Sharing Agreement.
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable.
   (d) Information about consent from data providers/curators. Yes. We have signed the Data Sharing Agreement for No.MS as explained before. CIFAR10 is publicly available.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable.
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Appendices

A BROADER IMPACT **18**

B LIMITATIONS **18**

C OTHER NOTATIONS **19**

D TRADITIONAL BOUNDS FAIL TO CAPTURE THE TRADE-OFF **19**

D.1 Choice of Optimum Regularization in the Traditional Setting 20

D.2 Alternative Formalization: Analysis of Ansatz 21

E ASYMPTOTIC EFFECT OF SYNTHETIC DATA IN KERNEL REGRESSION **22**

F TECHNICAL PROOFS OF MODIFIED KERNEL REGRESSION **23**

F.1 Proof of Lemma 2.1 23

F.2 Proof of Theorem 2.2 24

F.3 Proof of Corollary 2.2.1 26

G GENERALIZATION BOUND WITH MIXED REAL AND SYNTHETIC **27**

G.1 Lemmata 27

G.2 Stability of the Mixed Risk Minimizer 28

G.3 Proof of Theorem 3.1 31

H THEORETICAL RESULTS AND DISCUSSIONS OF DOMAIN SHIFT **33**

H.1 Proof of Theorem 5.1 33

H.2 Proof of Theorem 5.2 33

I EXPERIMENTAL SETUP **35**

I.1 Optimal Regularization in Kernel Ridge Regression 35

I.2 Natural Images on CIFAR-10 37

I.2.1 Experimental Details for CIFAR-10 37

I.3 Real-World Medical Imaging 39

I.3.1 Additional Results 39

I.4 Practical Insights 41

J EXPANDED RELATED WORK **43**

K NOTES ON THE CHOICE OF DISTANCE METRIC **45**

A BROADER IMPACT

This work shows how synthetic data can be effectively integrated with real data to improve the performance and generalization of downstream tasks in both in-domain and out-of-domain settings.

There are several potential benefits of our work:

1. Our framework enables practitioners to use an effective synthetic-to-real data ratio that yields improved performance at a reduced computational cost, therefore reducing carbon footprint.

2. Although our experiments focus on lesion segmentation, the underlying theory and insights are broadly applicable. Practitioners in various domains can leverage our framework to address challenges related to low-data regimes and domain shifts by exploiting powerful generative models to synthesize data, issues that are common across many applied fields.

3. We identify key factors necessary for evaluating the impact of synthetic data. This is particularly relevant in the current landscape, where a wide range of generative and foundation models are available to generate synthetic data. Our findings can help the community make more informed decisions about incorporating the generated samples from these models, particularly their quantity and quality.

4. Our results highlight the importance of distributional shift in achieving better performance, which in turn underscores the potential value of incorporating human feedback into the synthetic data generation process.

5. In scenarios involving biased datasets—closely related to our distribution shift setup—our framework offers a principled way to generate an adequate number of synthetic samples to improve model performance. This is particularly useful not only in data-scarce domains such as healthcare but also in datasets lacking diversity.

We also acknowledge potential risks and undesirable consequences associated with our approach. In efforts to maximize downstream task performance, practitioners may be incentivized to collect additional data or train more powerful generative models. This introduces several challenges:

1. Collecting extensive data about a subject raises concerns about responsible data acquisition.

2. Training larger generative models requires increased computational resources, which may have a greater environmental impact.

B LIMITATIONS

While our work provides theoretical insights and practical guidelines for combining synthetic and real data, several limitations remain:

1. Our analysis involves certain approximations to key parameters that affect the generalization bound and the optimal synthetic-to-real data ratio. The sensitivity of the results to our approach, and studying other ways of approximating them needs further investigation.

2. Although our theory aligns with empirical trends observed in lesion segmentation, we have not validated the proposed bounds across a broader range of applications. Extending the empirical evaluation to diverse domains would help assess the generality of our framework.

3. We focus on providing theoretical and practical insights and do not present a concrete algorithm that integrates a specific real dataset with a synthetic data generator. Developing such an algorithm would facilitate adoption in real-world settings.

4. Our experiments are restricted to the image modality. Investigating how the framework extends to other data types, such as text, audio, or multimodal settings, remains an open and promising direction for future work.

C OTHER NOTATIONS

We denote scalar or vector-valued Random Variable (RV) by $\mathbf{x}$, and collections of RVs by $\mathbf{X}$, with corresponding probability densities $p_{\mathbf{x}}$ and $p_{\mathbf{X}}$. Realizations of these variables are denoted by $\boldsymbol{x}$ and $\boldsymbol{X}$, respectively, with $\boldsymbol{x}$ taking values in a measurable space $\mathcal{X}$. The conditional distribution of a random variable $\mathbf{y}$ given $\mathbf{x} = \boldsymbol{x}$ is denoted by $p_{\mathbf{y}|\mathbf{x}=\boldsymbol{x}}$. The expectation of a measurable function $f : \mathcal{X} \to \mathbb{R}$ is written as $\mathbb{E}[f(\mathbf{x})] = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathbf{x}}}[f(\boldsymbol{x})]$. For integers $a \leq c \leq b$, we denote by $\mathbf{X}_{a:b} = \{\mathbf{x}_a, \mathbf{x}_{a+1}, \dots, \mathbf{x}_b\}$ a finite collection of RVs, and by $\mathbf{X}_{a:b}^{(\neq c)} = \mathbf{X}_{a:b} \setminus \{\mathbf{x}_c\}$ the subset excluding $\mathbf{x}_c$. The density $p_{\mathbf{X}_{a:b}}$ denotes the joint density of the variables in $\mathbf{X}_{a:b}$.

We use $[K]$ to denote the index set $\{1, \dots, K\}$, and reserve Latin letters for samples and Greek letters for parameters or distributions.

Definition C.1 (Lipschitz Continuity)**.** *A function $f : \mathcal{Z} \to \mathbb{R}^q$, for $(\mathcal{Z}, d_{\mathcal{Z}})$ a metric space, is ξ-Lipschitz if for all $z, z' \in \mathcal{Z}$, $\|f(z) - f(z')\| \leq \xi\, d_{\mathcal{Z}}(z, z')$.*

Let $\mathcal{R}_\lambda(h) = \mathbb{E}_{\mathbf{S}}[\mathcal{R}_\lambda(h, \mathbf{S})]$ denote the expected mixed loss, and $r_\lambda(h)$ the hybrid population risk:

$$r_\lambda(h_{\mathbf{S}}) = (1-\lambda) r(h) + \lambda \mathbb{E}_{\mathbf{x}\sim p'_{\mathbf{x}}}\left[\ell(h, \mathbf{x})\right]. \tag{3}$$

Definition C.2 (Generalization Gap)**.** *The* generalization error *of a hypothesis h is defined as the absolute difference between its population and empirical risks:*

$$g_{\mathbf{S}}(h) = \left|\mathcal{L}_{\mathcal{X}}(h) - \mathcal{L}_{\mathbf{S}}(h)\right|.$$

The generalization gap *of a learning algorithm is the expected generalization error:*

$$\mathcal{G} = \mathbb{E}_{p_{\mathrm{h},\mathbf{S}}}[g_{\mathbf{S}}(\mathrm{h})] = \mathbb{E}_{p_{\mathrm{h},\mathbf{S}}}\left[\left|\mathcal{L}_{\mathcal{X}}(\mathrm{h}) - \mathcal{L}_{\mathbf{S}}(\mathrm{h})\right|\right].$$

We can now define the generalization gap in the mixed-data setting as:

$$\mathcal{G} = \mathbb{E}_{p_{\mathrm{h},\mathbf{S}}}[g_{\mathbf{S}}(\mathrm{h})] = \mathbb{E}_{p_{\mathrm{h},\mathbf{S}}}\left[\left|r(\mathrm{h}) - \mathcal{R}_\lambda(\mathrm{h}, \mathbf{S})\right|\right].$$

D TRADITIONAL BOUNDS FAIL TO CAPTURE THE TRADE-OFF

We use the traditional bounds in learning theory to study the mixture of real and synthetic data. We assume a bounded loss $\ell \in [0, 1]$ and hypothesis class $\mathcal{H}$. The goal is to bound the distance between $\hat{\mathcal{R}}_{p_{\mathbf{x}}}(h) = \mathcal{L}_{\mathbf{S}}(h)$ and $\mathcal{R}_{p_{\mathbf{x}}}(h) = r(h)$ as expressed in Section 3.

Bias via Mixture Mismatch For any $f \in \mathcal{H}$, and $\alpha \in [0, 1]$, let $q_\alpha := (1-\alpha)p_{\mathbf{x}} + \alpha p'_{\mathbf{x}}$. We have

$$\begin{aligned}\mathcal{R}_{p_{\mathbf{x}}}(f) &= R_{q_\alpha}(f) + \alpha\big(\mathcal{R}_{p_{\mathbf{x}}}(f) - \mathcal{R}_{p'_{\mathbf{x}}}(f)\big) && (4)\\ &\leq \mathcal{R}_{q_\alpha}(f) + \alpha\, \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}), && (5)\end{aligned}$$

where $\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}) := \sup_{f\in\mathcal{H}} \left|\mathbb{E}_{p_{\mathbf{x}}}[\ell(f)] - \mathbb{E}_{p'_{\mathbf{x}}}[\ell(f)]\right|$.

Estimation for $\mathcal{R}_{q_\alpha}(f)$ Let $S_N \sim p_{\mathbf{x}}^{\otimes N}$ and $S'_M \sim p_{\mathbf{x}}'^{\otimes M}$ be independent samples of size N, M, from $p_{\mathbf{x}}$ and $p'_{\mathbf{x}}$, respectively. Suppose we train on a mixture of these samples so the empirical risk becomes

$$\hat{\mathcal{R}}_{q_\alpha}(f) = (1-\alpha)\,\hat{\mathcal{R}}_{S_N}(f) + \alpha\,\hat{\mathcal{R}}_{S'_M}(f),$$

and the population risk is $\mathcal{R}_{q_\alpha}(f)$. By standard uniform convergence bounds (Shalev-Shwartz et al., 2010), with probability at least $1-\delta$, simultaneously for all $f \in \mathcal{H}$,

$$\begin{aligned}\mathcal{R}_{p_{\mathbf{x}}}(f) &\leq \hat{\mathcal{R}}_{S_N}(f) + \varepsilon_N,\\ \mathcal{R}_{p'_{\mathbf{x}}}(f) &\leq \hat{\mathcal{R}}_{S'_M}(f) + \varepsilon_M,\end{aligned}$$

where, for example,

$$\varepsilon_N = 2\,\mathfrak{R}_N^{(p_{\mathbf{x}})}(\ell \circ \mathcal{H}) + \sqrt{\tfrac{\log(4/\delta)}{2N}},$$
$$\varepsilon_M = 2\,\mathfrak{R}_M^{(p'_{\mathbf{x}})}(\ell \circ \mathcal{H}) + \sqrt{\tfrac{\log(4/\delta)}{2M}}.$$

Here $\mathfrak{R}_n^{(P)}$ is the Rademacher complexity under distribution P. Now, multiplying the two bounds by $(1-\alpha)$ and α, respectively and adding, we have

$$\mathcal{R}_{q_\alpha}(f) \;\le\; \hat{\mathcal{R}}_{q_\alpha}(f) + (1-\alpha)\varepsilon_N + \alpha\varepsilon_M. \tag{6}$$

Combine with bias Let $\hat{f}_\alpha \in \arg\min_{f\in\mathcal{H}} \hat{\mathcal{R}}_{q_\alpha}(f)$. Then, with probability at least $1-\delta$, by combining Equations 5 and 6, we have:

$$\mathcal{R}_{p_{\mathbf{x}}}(\hat{f}_\alpha) \;\le\; \hat{\mathcal{R}}_{q_\alpha}(\hat{f}_\alpha) + (1-\alpha)\varepsilon_N + \alpha\varepsilon_M + \alpha\,\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}})\,. \tag{7}$$

Suppose a favourable regime where the Rademacher complexity shrinks at the canonical rate $1/\sqrt{n}$ with a distribution and sample size independent constant. Note that kernel methods with bounded kernels (Section 2) and Lipschitz losses (Section 3) are examples of such regimes. Therefore, we can bound the Rademacher complexities as

$$\mathfrak{R}_n^{(P)}(\ell \circ \mathcal{H}) \le C_{\ell\circ\mathcal{H}}/\sqrt{n}$$

for some constant $C_{\ell\circ\mathcal{H}}$ independent of P and n. Then we can simplify the above to

$$\varepsilon_N \le 2C_{\ell\circ\mathcal{H}}/\sqrt{N} + \sqrt{\log(4/\delta)/(2N)}, \qquad \varepsilon_M \le 2C_{\ell\circ\mathcal{H}}/\sqrt{M} + \sqrt{\log(4/\delta)/(2M)}.$$

Therefore, there exists a constant C_δ such that the following holds:

$$\varepsilon_N \le \frac{C_\delta}{\sqrt{N}}, \qquad \varepsilon_M \le \frac{C_\delta}{\sqrt{M}}$$

D.1 Choice of Optimum Regularization in the Traditional Setting

In our approach, we use an empirical risk of the form

$$\tilde{\mathcal{R}}(f) = \frac{1}{N+M}\sum_{i=1}^{N} \ell(f(X_i), Y_i) + \frac{1}{N+M}\sum_{j=1}^{M} \ell(f(X'_i), Y'_i)\,, \tag{8}$$

where $S_N = ((X_i, Y_i); i \in [N])$ and $S'_M = \big((X'_j, Y'_j); j \in [M]\big)$ are samples from $p_{\mathbf{x}}$ and $p'_{\mathbf{x}}$, respectively.

To connect to the previous analysis, we can choose $\alpha_N = \frac{M}{N+M}$. With this choice of α, and by Equation 7, we have

$$\mathcal{R}_D(\hat{f}_{\alpha_N}) \;\le\; \hat{\mathcal{R}}_{q_{\alpha_N}}(\hat{f}_{\alpha_N}) + \frac{N}{M+N}\frac{C_\delta}{\sqrt{N}} + \frac{M}{M+N}\frac{C_\delta}{\sqrt{M}} + \frac{M}{N+M}\,\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}).$$

Assume that $M = \lambda N$ for some $\lambda > 0$. Then, we have

$$\mathcal{R}_D(\hat{f}_{\alpha_N}) \;\le\; \hat{\mathcal{R}}_{q_{\alpha_N}}(\hat{f}_{\alpha_N}) + C_\delta\frac{1+\sqrt{\lambda}}{(1+\lambda)\sqrt{N}} + \frac{\lambda}{1+\lambda}\,\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}).$$

Note that the second term, $C_\delta\frac{1+\sqrt{\lambda}}{(1+\lambda)\sqrt{N}}$, which arises from the variance component, attains a non-trivial maximum. When $\lambda = 0$, the bound reduces to the error using only real data (i.e., without synthetic augmentation). As λ increases, the variance term eventually decreases, while the bias term increases but only up to the discrepancy between the two distributions. Consequently, the optimal λ is either zero or infinity, depending on whether the discrepancy is smaller than the variance term at $\lambda = 0$. Under a fixed computational budget, however, this yields a non-trivial optimal choice of λ.

Optimal Choice of α (Effective Sample View) Let us bound the estimation error by

$$(1-\alpha)\varepsilon_N + \alpha\varepsilon_M \lesssim \frac{c}{\sqrt{N_\alpha}}, \qquad N_\alpha = (1-\alpha)N + \alpha M,$$

with $c \approx \tilde{2}C_{\ell\circ\mathcal{H}} + \sqrt{\log(1/\delta)}$. Thus, we need to minimize the following bound:

$$\rho(\alpha) = \frac{c}{\sqrt{(1-\alpha)N + \alpha M}} + \alpha \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}).$$

Let $\Delta = M - N$ and $n^\star = \left(\frac{c\,\Delta}{2\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}})}\right)^{2/3}$. Then, the optimal mixing weight is

$$\alpha^\star = \mathrm{clip}_{[0,1]}\left(\frac{n^\star - N}{M - N}\right).$$

This provides us with the following decision rule:

$$\alpha^\star = \begin{cases} 0, & \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}) \geq \frac{c}{\sqrt{N}}, \\ 1, & \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}) \leq \frac{c}{\sqrt{M}}, \\ \frac{n^\star - N}{M-N}, & \text{otherwise.} \end{cases}$$

Optimal M with Fixed N (Uniform Concatenation) If training is done on the mixture dataset, following Equation 8 with each point weighted equally, i.e. ,

$$\alpha = \frac{M}{N+M} = \frac{\lambda}{1+\lambda}, \qquad M = \lambda N,$$

then, we need to optimize with respect to the following bound:

$$\rho(\lambda) = \frac{c}{\sqrt{N(1+\lambda)}} + \frac{\lambda}{1+\lambda}\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}), \qquad \lambda \geq 0.$$

This formalization gives us the following threshold behavior:

$$M^\star = \begin{cases} 0, & \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}) \;\geq\; \frac{c}{\sqrt{N}}, \\ \text{“as large as allowed,”} & \mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}}) \;<\; \frac{c}{\sqrt{N}}. \end{cases}$$

A finite *balance point* is obtained by equating variance and bias:

$$M_{\mathrm{bal}} = \left(\left(\frac{2c}{\mathrm{IPM}_{\ell\circ\mathcal{H}}(p_{\mathbf{x}}, p'_{\mathbf{x}})\sqrt{N}}\right)^2 - 1\right)_+ N.$$

Therefore, although traditional bounds offer an intuitive understanding of the ERM's behavior, they do not capture its complete behavior and can be overly loose.

To solve this problem and bridge the gap between the generalization bounds and what we observe in practice, we propose an alternative option for the optimization problem that we introduce in the next section.

D.2 Alternative Formalization: Analysis of Ansatz

In practice, we have the following ERM:

$$f_N = \arg\min_{f\in\mathcal{H}_k} \sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \sum_{m=1}^{M} (\tilde{\mathrm{y}}_m - f(\tilde{\mathrm{x}}_m))^2. \tag{9}$$

Notice that the objective is equivalent to

$$\begin{aligned}
&\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \sum_{m=1}^{M} (\tilde{\mathrm{y}}_m - f(\tilde{\mathrm{x}}_m))^2 \\
&\overset{\circ}{=} \frac{1}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \frac{1}{N}\sum_{m=1}^{M} (\tilde{\mathrm{y}}_m - f(\tilde{\mathrm{x}}_m))^2 \\
&\overset{\circ}{=} \frac{1}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \frac{M}{N}\frac{1}{M}\sum_{m=1}^{M} (\tilde{\mathrm{y}}_m - f(\tilde{\mathrm{x}}_m))^2 \ .
\end{aligned}$$

We can now approximate the empirical loss over the synthetic samples with the corresponding population loss:

$$\approx \frac{1}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \frac{M}{N}\|f-g\|^2 \,.$$

Letting $\lambda = M/N$:

$$= \frac{1}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \lambda\|f-g\|^2 \,.$$

Using $\lambda = \tilde{\lambda}/(1-\tilde{\lambda})$, we have

$$\begin{aligned}
&= \frac{1}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \frac{\tilde{\lambda}}{1-\tilde{\lambda}}\|f-g\|^2 \\
&\overset{\circ}{=} \frac{1-\tilde{\lambda}}{N}\sum_{n=1}^{N} (\mathrm{y}_n - f(\mathrm{x}_n))^2 + \tilde{\lambda}\|f-g\|^2.
\end{aligned}$$

In this formalization, we deliberately omit the variance term arising from the synthetic samples, as it is expected to play only a secondary role and may obscure the regularizing effect introduced by the synthetic data. As we see in Sections 2 and 3, this formalization helps us to improve the bound for the U-shape observed in practice, as opposed to the classical case, where methods fail to capture this trade-off.

E ASYMPTOTIC EFFECT OF SYNTHETIC DATA IN KERNEL REGRESSION

Suppose we have M synthetic samples $\{(\tilde{x}_m, \tilde{y}_m)\}_{m=1}^{M}$, where $\tilde{x}_m \sim p(x)$ i.i.d., and $\tilde{y}_m = g(\tilde{x}_m)$. We assume these synthetic samples are noiseless, reflecting access to the exact synthetic data generator. Then the ERM objective

$$f_N = \operatorname*{arg\,min}_{f\in\mathcal{H}_k} \ \frac{1}{N}\sum_{n=1}^{N}(y_n - f(x_n))^2 + \lambda\,\frac{1}{M}\sum_{m=1}^{M}\big(f(\tilde{x}_m) - g(\tilde{x}_m)\big)^2$$

satisfies, by the (strong) law of large numbers,

$$\lim_{M\to\infty}\frac{1}{M}\sum_{m=1}^{M}\big(f(\tilde{x}_m) - g(\tilde{x}_m)\big)^2 = \mathbb{E}_{\tilde{x}\sim p}\big[(f(\tilde{x}) - g(\tilde{x}))^2\big] = \big\langle f-g,\, T_K(f-g)\big\rangle_{\mathcal{H}_k},$$

where the kernel integral operator $T_K : \mathcal{H}_k \to \mathcal{H}_k$ is defined by

$$(T_K\, h)(\cdot) = \int K(\cdot, x)\, h(x)\, p(x)\, \mathrm{d}x.$$

Note that while the synthetic covariates $\tilde{x}_m$ are drawn i.i.d. from the same marginal distribution as the real data, the synthetic labels $\tilde{y}_m$ follow a potentially different mapping g, as determined by the data generator.

Equivalence of $L^2(p)$ and RKHS norms By Mercer's theorem (Mercer, 1909), the operator T_K admits the spectral decomposition

$$K(x,y) = \sum_{i=1}^{\infty} \mu_i \, \phi_i(x) \, \phi_i(y),$$

where $\{\phi_i\}$ form an orthonormal basis in $L^2(p)$ and $\mu_i > 0$ are the eigenvalues. Any function $h \in \mathcal{H}_k$ can be written as $h = \sum_i a_i \sqrt{\mu_i} \, \phi_i$, yielding

$$\|h\|^2_{L^2(p)} = \sum_{i=1}^{\infty} \mu_i \, a_i^2, \qquad \|h\|^2_{\mathcal{H}_k} = \sum_{i=1}^{\infty} a_i^2.$$

If the nonzero eigenvalues satisfy $0 < \mu_{\min} \le \mu_i \le \mu_{\max} < \infty$, then

$$\mu_{\min} \|h\|^2_{\mathcal{H}_k} \le \|h\|^2_{L^2(p)} \le \mu_{\max} \|h\|^2_{\mathcal{H}_k},$$

so the norms are equivalent up to constants:

$$\|h\|^2_{L^2(p)} \asymp \|h\|^2_{\mathcal{H}_k}.$$

Under this spectral assumption, the $L^2(p)$ term $\mathbb{E}_{\tilde{x}\sim p}[(f(\tilde{x}) - g(\tilde{x}))^2]$ appearing in the infinite-M limit is thus proportional to $\|f - g\|^2_{\mathcal{H}_k}$.

Note that we use this equivalence to motivate our setup, we study the effect of limited synthetic data in Section 3 more precisely.

F TECHNICAL PROOFS OF MODIFIED KERNEL REGRESSION

F.1 Proof of Lemma 2.1

Lemma 2.1. *Let $K_N \in \mathbb{R}^{N\times N}$ be the empirical kernel matrix with entries $(K_N)_{ij} = K(\mathrm{x}_i, \mathrm{x}_j)$. Define the integral operator $T_K : L^2(p_x) \to L^2(p_x)$ by $(T_K f)(\mathrm{x}) = \int K(\mathrm{x}, x') f(x') \, dp_x(x') = \mathbb{E}_{\mathrm{x}'}\left[K(\mathrm{x}, \mathrm{x}') f(\mathrm{x}')\right]$. Let $\lambda_N = N\lambda$. Then the solution to Equation 1 has the closed-form representation:*

$$\boldsymbol{\alpha} = \left(K_N + \lambda_N I\right)^{-1} \left(K_N \boldsymbol{\alpha}_\star + \lambda_N \boldsymbol{\beta} + \boldsymbol{\varepsilon}\right) ,$$

where $\boldsymbol{\alpha}_\star$, $\boldsymbol{\beta}$, and $\boldsymbol{\varepsilon}$ are the coefficients of $f_\star$, g, and the noise vector in the training basis.

Proof. Let us define $\lambda = \frac{\tilde{\lambda}}{1-\tilde{\lambda}}$. We can rewrite the ERM using the Representer theorem as following:

$$\boldsymbol{\alpha} = \arg\min_{\hat{\boldsymbol{\alpha}}} \frac{1}{N} \|\mathbf{y} - K_N \hat{\boldsymbol{\alpha}}\|^2 + \lambda \|\hat{\boldsymbol{\alpha}} - \boldsymbol{\beta}\|^2_{\mathcal{H}_k} + \lambda \|g_\perp\|^2_{\mathcal{H}_k} .$$

Finite-sample solution. Taking the derivation with respect to $\boldsymbol{\alpha}$, we have

$$K_N \left[(K_N + \lambda_N I)\boldsymbol{\alpha} - \mathbf{y} - \lambda_N \boldsymbol{\beta}\right] = 0 \tag{10}$$

Solving the optimization, similarly to the standard regularized kernel regression, we achieve:

$$\boldsymbol{\alpha} = \left(K_N + \lambda_N I\right)^{-1} \left(\mathbf{y} + \lambda_N \boldsymbol{\beta}\right),$$

where we conclude the proof by noting that $\mathbf{y} = K_N \boldsymbol{\alpha}_\star + \boldsymbol{\varepsilon}$. This also results in the fact that $f_N(\mathrm{x}) = K_{\mathrm{x}} \left(K_N + \lambda_N I\right)^{-1} \left(\mathbf{y} + \lambda_N \boldsymbol{\beta}\right)$, where $K_{\mathrm{x}} = (K(\mathrm{x}, \mathrm{x}_1), \ldots, K(\mathrm{x}, \mathrm{x}_N))$.

We now study the behavior of this closed-form solution in the population limit, which becomes useful later.

Population limit and expectation. We use Equation 10, so we have:

$$\begin{aligned} K_N \left[(K_N + \lambda_N I)(\boldsymbol{\alpha} - \boldsymbol{\beta}) - \mathbf{y} + K_N \boldsymbol{\beta}\right] &= 0 \\ \boldsymbol{\alpha} - \boldsymbol{\beta} &= \left(K_N + \lambda_N I\right)^{-1} \left(\mathbf{y} - K_N \boldsymbol{\beta}\right) \\ f_N(\mathrm{x}) - g(\mathrm{x}) &= K_{\mathrm{x}} \left(K_N + \lambda_N I\right)^{-1} \left(\mathbf{y} - K_N \boldsymbol{\beta}\right). \end{aligned}$$

Now, consider the population limit where the sample size $N \to \infty$. The empirical kernel matrix K_N converges to the integral operator T_K, which is a classic approach in kernel ridge regression (Singh and Vijaykumar, 2023). Therefore, $T_K \approx \frac{1}{N}\sum_{n=1}^{N} K(\cdot, \mathrm{x}_n) \otimes K(\cdot, \mathrm{x}_n)$, which means $K_N \approx NT_K$. So, we have

$$f_N - g = (T_K + \lambda I)^{-1}\left(T_K f_\star - T_K g + \frac{1}{N}\boldsymbol{\varepsilon}\right)$$
$$\mathbb{E}_\varepsilon\left[f_N - g\right] = (T_K + \lambda I)^{-1} T_K (f_\star - g)\,.$$

Noting that $\mathbb{E}_\varepsilon[g] = g$, we get the following result in the population limit:

$$\mathbb{E}_\varepsilon\left[f_N\right] = g + (T_K + \lambda I)^{-1} T_K (f_\star - g)\,. \tag{11}$$

□

F.2 Proof of Theorem 2.2

Theorem 2.2. *Under Assumption 2.1, for the kernel regression problem defined in Equation 1 and any fixed regularization parameter $\lambda > 0$, the test error admits the bound:*

$$\mathcal{R}_N(\lambda; g) = \mathcal{O}\left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\lambda^2} + \lambda^{2-\frac{1}{4r}}\mathcal{D}(f_\star, g)\right)\,,$$

where $\mathcal{D}(f_\star, g)^2 = \sum_{j=1}^{\infty} \frac{1}{\mu_j^2}(\theta_j - \omega_j)^2$ denotes the discrepancy between the target function $f_\star$ and the synthetic generator g.

Proof. To bound the test error, we use the bias-variance decomposition in Definition 2.1. We start with the variance term. We follow the approach of Misiakiewicz and Saeed (2024). So, we have:

$$\mathcal{V} = \mathbb{E}_{\mathbf{x},\varepsilon}\left[(f_N(\mathbf{x}) - \mathbb{E}_\varepsilon\left[f_N(\mathbf{x})\right])^2\right] \tag{12}$$
$$= \mathbb{E}_{\mathbf{x},\varepsilon}\left[\left(\frac{1}{N}K_{\mathbf{x}}\left(T_K + \lambda I\right)^{-1}\boldsymbol{\varepsilon}\right)\right] \tag{13}$$
$$= \frac{\sigma^2}{N^2}\mathrm{tr}\left(NT_K^2\left(T_K + \lambda I\right)^{-2}\right) \tag{14}$$
$$= \frac{\sigma^2}{N}\sum_{j=1}^{\infty}\frac{\mu_j^2}{(\mu_j + \lambda)^2}\,. \tag{15}$$

Note that the above discussion assumes the population limit. An analogous behaviour holds in the finite-sample setting. Define $\kappa = \sup_{\boldsymbol{x}\in\mathcal{X}}\sqrt{K(\boldsymbol{x}, \boldsymbol{x})}$. Then

$$\mathcal{V} = \mathbb{E}_{\mathbf{x},\varepsilon}\left[(f_N(\mathbf{x}) - \mathbb{E}_\varepsilon[f_N(\mathbf{x})])^2\right] \tag{16}$$
$$= \mathbb{E}_{\mathbf{x},\varepsilon}\left[\left(\tfrac{1}{N}K_{\mathbf{x}}(K_N + \lambda I)^{-1}\boldsymbol{\varepsilon}\right)^2\right] \tag{17}$$
$$\leq \frac{\kappa^2\sigma^2}{N\lambda^2}\,, \tag{18}$$

where the inequality follows from $\|(K_N + \lambda I)^{-1}\| \leq \frac{1}{\lambda}$ and the bound $\|K_{\mathbf{x}}\|^2 \leq N\kappa^2$. Now, let us bound the bias term. We have:

$$\mathcal{B}^2 = \mathbb{E}_{\mathbf{x}}\left[(f_\star - \mathbb{E}_\varepsilon\left[f_N\right])(\mathrm{x})\right]^2$$
$$\leq \mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_\lambda\|_{\mathcal{H}_K}^2\right] + \|f_\star - f_\lambda\|_{p_{\mathbf{x}}}^2\,,$$

where the last line is resulted from Jensen's inequality and triangle inequality. Moreover, note that $\|f_\star - f_\lambda\|_{p_{\mathbf{x}}}^2 \leq \kappa^2\|f_\star - f_\lambda\|_{\mathcal{H}}^2$. We define f_λ as the population limit of f_N:

$$f_\lambda = g + (T_K + \lambda I)^{-1}T_K(f_\star - g)\,. \tag{19}$$

To bound the bias term, we also define an additional auxiliary function $f_{N,\lambda}$:

$$f_{N,\lambda} = g + (K_N + \lambda I)^{-1} T_K(f_\star - g)\,.$$

This function helps us compute the first term of bias.

Population and sampling bias $\mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_\lambda\|^2_{\mathcal{H}_K}\right]$. We rewrite this term as follows:

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_\lambda\|^2_{\mathcal{H}_K}\right] &\leq \mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_{N,\lambda}\|^2_{\mathcal{H}_K}\right] + \mathbb{E}_{\mathbf{S}}\left[\|f_{N,\lambda} - f_\lambda\|^2_{\mathcal{H}_K}\right] \\
&\leq \mathbb{E}_{\mathbf{S}}\left[\|(K_N + \lambda I)^{-1}(K_N - T_K)(f_\star - g)\|^2_{\mathcal{H}_K}\right] + \mathbb{E}_{\mathbf{S}}\left[\|f_{N,\lambda} - f_\lambda\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{1}{\lambda^2} E_{\mathbf{S}}\left[\|(K_N - T_K)(f_\star - g)\|^2_{\mathcal{H}_K}\right] + \mathbb{E}_{\mathbf{S}}\left[\|f_{N,\lambda} - f_\lambda\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \mathbb{E}_{\mathbf{S}}\left[\|f_{N,\lambda} - f_\lambda\|^2_{\mathcal{H}_K}\right]\,,
\end{aligned}$$

where we used the fact that $\|(K_N + \lambda I)^{-1}\| \leq \frac{1}{\lambda}$, and the last inequality is proved in Smale and Zhou (2005)[Theorem 3]. We continue the bound by first noticing that Equation 19 gives us $(T_K + \lambda I)(f_\lambda - g) = T_K(f_\star - g)$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_\lambda\|^2_{\mathcal{H}_K}\right] &\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \mathbb{E}_{\mathbf{S}}\left[\|f_{N,\lambda} - f_\lambda\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \mathbb{E}_{\mathbf{S}}\left[\|(K_N + \lambda I)^{-1}(T_K + \lambda I)(f_\lambda - g) - (f_\lambda - g)\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \mathbb{E}_{\mathbf{S}}\left[\|(K_N + \lambda I)^{-1}(T_K - K_N)(f_\lambda - g)\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \frac{1}{\lambda^2}\mathbb{E}_{\mathbf{S}}\left[\|(T_K - K_N)(f_\lambda - g)\|^2_{\mathcal{H}_K}\right] \\
&\leq \frac{\kappa^2\|f_\star - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2} + \frac{\kappa^2\|f_\lambda - g\|^2_{p_{\mathbf{x}}}}{N\lambda^2}\,,
\end{aligned}$$

where we have used Smale and Zhou (2005)[Theorem 3] once more. Moreover, we note that Equation 19 is also the solution to the following Kernel optimization:

$$f_\lambda = g + \underset{f\in\mathcal{H}_K}{\arg\min}\left\{\|f - (f_\star - g)\|^2_{p_{\mathbf{x}}} + \lambda\|f\|^2_{\mathcal{H}_K}\right\}\,.$$

Therefore, setting f to zero, we have $\|f_\lambda - (f_\star - g)\|^2_{p_{\mathbf{x}}} + \lambda\|f_\lambda\|^2_{\mathcal{H}_K} \leq \|f_\star - g\|^2_{p_{\mathbf{x}}}$, from which we can conclude that $\|f_\lambda - g\|^2_{p_{\mathbf{x}}} \leq 2\|f_\star - g\|^2_{p_{\mathbf{x}}}$. Putting all these results together and the fact that $\|f_\star - g\|^2_{p_{\mathbf{x}}} \leq \sup_{\boldsymbol{x}} K(x,x)\|f_\star - g\|^2_{\mathcal{H}_K}$, we have:

$$\mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon\left[f_N\right] - f_\lambda\|^2_{\mathcal{H}_K}\right] \leq \frac{3\kappa^4\|f_\star - g\|^2_{\mathcal{H}}}{N\lambda^2}\,. \tag{20}$$

Population bias $\|f_\star - f_\lambda\|^2_{\mathcal{H}}$. To bound this term, we substitute the Mercer decomposition of $f_\star$ and g, and the fact that the eigenvalues of $(T_K + \lambda I)^{-1}$ are $1/(\lambda + \mu_j)$ as following:

$$\begin{aligned}
f_\lambda &= g + \left(T_K + \lambda I\right)^{-1} T_K(f_\star - g) \\
&= \sum_j^\infty \left(\frac{\mu_j}{\mu_j + \lambda}\theta_j + \frac{\lambda}{\mu_j + \lambda}\omega_j\right)\phi_j\,.
\end{aligned}$$

Therefore, we have

$$\|f_\star - f_\lambda\|^2_{\mathcal{H}} = \|\sum_{j=1}^\infty \frac{\lambda}{\mu_j + \lambda}(\theta_j - \omega_j)\phi_j\|^2\,,$$

We can bound this bias term as follows by noting that $\{\phi_j\}_j$ consist the orthonormal basis:

$$\|f_\star - f_\lambda\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} \frac{\lambda^2}{(\mu_j+\lambda)^2}(\theta_j - \omega_j)^2 \,. \tag{21}$$

Now, combining the results of Equations 20 and 21 gives us an upper-bound for $\mathcal{B}^2$, and combining them with Equation 18 gets a bound for the test error. We have:

$$\begin{aligned}
\mathcal{R}_N(\lambda; g) &= \frac{3\kappa^4\|f_\star - g\|_{\mathcal{H}}^2}{N\lambda^2} + \sum_{j=1}^{\infty} \frac{\lambda^2}{(\mu_j+\lambda)^2}(\theta_j - \omega_j)^2 + \frac{\kappa^2\sigma^2}{N\lambda^2} \\
&\leq \frac{3\kappa^4\|f_\star - g\|_{\mathcal{H}}^2}{N\lambda^2} + \lambda^2 \sqrt{\left(\sum_{j=1}^{\infty} \frac{\mu_j^2}{(\mu_j+\lambda)^2}\right)\left(\sum_{j=1}^{\infty} \frac{1}{\mu_j^2}(\theta_j - \omega_j)^2\right)} + \frac{\kappa^2\sigma^2}{N\lambda^2} \,,
\end{aligned}$$

where the inequality is due to the Cauchy-Schwarz inequality. Since $\sum_{j=1}^{\infty} \frac{1}{\mu_j^2}(\theta_j - \omega_j)^2 = \mathcal{D}(f_\star, g)^2$, we can now write:

$$\begin{aligned}
\mathcal{R}_N(\lambda; g) &= \mathcal{O}\left(\frac{3\kappa^4\|f_\star - g\|_{\mathcal{H}}^2}{N\lambda^2} + \lambda^2 \sqrt{\mathcal{D}(f_\star, g)^2 \left(\sum_{j=1}^{\infty} \frac{\mu_j^2}{(\mu_j+\lambda)^2}\right)} + \frac{\kappa^2\sigma^2}{N\lambda^2}\right) \\
&= \mathcal{O}\left(\frac{3\kappa^4\|f_\star - g\|_{\mathcal{H}}^2}{N\lambda^2} + \lambda^2 \mathcal{D}(f_\star, g)\sqrt{\int_0^\infty \left(\frac{x^{-2r}}{x^{-2r}+\lambda}\right)^2 dx} + \frac{\kappa^2\sigma^2}{N\lambda^2}\right) \\
&= \mathcal{O}\left(\frac{3\kappa^4\|f_\star - g\|_{\mathcal{H}}^2}{N\lambda^2} + \lambda^{2-\frac{1}{4r}} \mathcal{D}(f_\star, g) C_r + \frac{\kappa^2\sigma^2}{N\lambda^2}\right) \,,
\end{aligned}$$

where

$$C_r = \left(\int_0^\infty \left(\frac{x^{-2r}}{x^{-2r}+\lambda}\right)^2 dx\right)^{1/2} = \left(\frac{1}{2r}\int_0^\infty \frac{v^{1-1/2r}}{(v+1)^2}\, dv\right)^{1/2} = \sqrt{\frac{B(1/2r, 2-1/2r)}{2r}} \,,$$

with $B(z_1, z_2)$ denoting the beta function. We conclude by noting that $\mathcal{D}(f_\star, g) \geq \|f_\star - g\|_{\mathcal{H}}^2$. □

F.3 Proof of Corollary 2.2.1

Corollary 2.2.1. *Under the assumptions of Theorem 2.2, the optimal regularization parameter that minimizes the test error is given by*

$$\lambda^\star \asymp \left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} \,.$$

Setting $\lambda = \frac{M}{N}$, the optimal number of synthetic samples satisfies:

$$M^\star \asymp \left(1 + \frac{\sigma^2}{\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} N^{\frac{12r+1}{16r+1}} \,.$$

Proof. The result follows by minimizing the bound in Theorem 2.2:

$$\mathcal{R}_N(\lambda; g) = \mathcal{O}\left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\lambda^2} + \lambda^{2-\frac{1}{4r}} \mathcal{D}(f_\star, g)\right) \,,$$

We differentiate the right-hand side with respect to λ and set the derivative to zero:

$$\frac{\partial \mathcal{R}_N}{\partial \lambda} = \left(2 - \frac{1}{4r}\right) \lambda^{1-\frac{1}{4r}} \mathcal{D}(f_\star, g) - 2\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N} \lambda^{-3} = 0 \,.$$

Solving for λ gives

$$\lambda^\star = \left(\frac{8r\left(\mathcal{D}(f_\star, g) + \sigma^2\right)}{(8r-1)N\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} \asymp \left(\frac{\mathcal{D}(f_\star, g) + \sigma^2}{N\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}}.$$

Substituting $\lambda^\star = \frac{M^\star}{N}$ yields

$$M^\star = N\lambda^\star \asymp \left(1 + \frac{\sigma^2}{\mathcal{D}(f_\star, g)}\right)^{\frac{4r}{16r+1}} N^{\frac{12r+1}{16r+1}},$$

completing the proof. □

G GENERALIZATION BOUND WITH MIXED REAL AND SYNTHETIC

G.1 Lemmata

Definition G.1. *The upper packing dimension of a measure ν is the quantity d^* defined by:*

$$d^* := \operatorname{ess\,sup}(\Phi^*), \quad \Phi^*(x) := \limsup_{\delta \to 0} \frac{\log p_\delta(x)}{\log \delta}.$$

Definition G.2 ($\mathcal{D}$-Regularity Clerico et al. (2022))**.** *Let $\mathcal{D}$ be a measurable map $\mathcal{P} \times \mathcal{P} \to [0, +\infty]$. Fix $\mu \in \mathcal{P}$ and $\xi \geq 0$. We say that a function $f : \mathcal{Z} \to \mathbb{R}$ is $R_{\mathcal{D}}(\xi)$-regular with respect to μ if $f \in L^1(\mu)$ and for every $\nu \in \mathcal{P}$ such that* $\mathrm{Supp}(\nu) \subseteq \mathrm{Supp}(\mu)$ *and $f \in L^1(\nu)$,*

$$|\mathbb{E}_\mu[f(Z)] - \mathbb{E}_\nu[f(Z)]| \leq \xi\, \mathcal{D}(\mu, \nu).$$

Lemma G.1. *Consider a mapping $A : \mathcal{S}_N \to \mathcal{H}$ and define the random variable $\tilde{\mathrm{x}} \sim p_{\mathrm{x}}$ such that $\tilde{\mathrm{x}} \perp\!\!\!\perp \mathbf{S}$. Suppose there exists $\varepsilon \geq 0$ such that for any $i \in \{1, ..., N\}$, it holds that*

$$\sup_i \mathbb{E}[\ell(h^i, \mathrm{x}_i) - \ell(h, \mathrm{x}_i)] \leq \varepsilon, \tag{22}$$

where $h = A(\mathbf{S}), h^i = A(\mathbf{S}^i)$ and $\mathbf{S}^i = \{\mathrm{x}_1, ..., \mathrm{x}_{i-1}, \tilde{\mathrm{x}}, \mathrm{x}_{i+1}, ..., \mathrm{x}_N\}$. Then it holds that

$$\mathbb{E}[\mathcal{L}_{\mathcal{X}}(A(\mathbf{S})) - \mathcal{L}_{\mathbf{S}}(A(\mathbf{S}))] \leq \varepsilon.$$

Proof. Follows from Lemma 7 of Bousquet and Elisseeff (2002). □

Now we obtain Wasserstein continuity bounds using a similar technique to Polyanskiy and Wu (2015); Raginsky et al. (2017).

Lemma G.2. *Suppose that Assumptions 3.1–3.3 hold, then for any $h \in \mathcal{H}$,*

$$\left|\mathbb{E}_{p_{\mathrm{x}}}[\ell(h, \mathrm{x})] - \mathbb{E}_{p'_{\mathrm{x}}}[\ell(h, \mathrm{x})]\right| \leq \xi\, \mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}), \qquad \xi := 3\sqrt{1 + L^2}\left(\|\nabla c(0, 0)\| + M\tau\right), \tag{23}$$

where $\tau^2 = 2B_{\mathcal{H}}^2 + (1 + 2L^2)(\sigma^2 + \sigma'^2)$, and $B_{\mathcal{H}} = \sup_{h \in \mathcal{H}} \|h(0)\|$.

Proof. Define the map $\phi_h(\mathrm{x}) = (h(\mathrm{x}), \mathrm{x})$. Since h is L-Lipschitz,

$$\|\phi_h(\mathrm{x}) - \phi_h(\mathrm{x}')\|^2 = \|h(\mathrm{x}) - h(\mathrm{x}')\|^2 + \|\mathrm{x} - \mathrm{x}'\|^2 \leq (1 + L^2)\|\mathrm{x} - \mathrm{x}'\|^2,$$

so ϕ_h is $\sqrt{1 + L^2}$-Lipschitz. It follows that

$$\mathcal{W}_2(\phi_h \# p_{\mathrm{x}},\, \phi_h \# p'_{\mathrm{x}}) \leq \sqrt{1 + L^2}\, \mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}). \tag{24}$$

Let (X, Y) be an optimal $\mathcal{W}_2$-coupling of p_{x} and p'_{x}, and set $Z = \phi_h(X)$ and $W = \phi_h(Y)$. By M-smoothness of c,

$$c(Z) - c(W) \leq \langle \nabla c(W),\, Z - W \rangle + \frac{M}{2}\|Z - W\|^2.$$

Taking expectations and applying Cauchy–Schwarz to the first term gives

$$\mathbb{E}[c(Z)] - \mathbb{E}[c(W)] \leq \sqrt{\mathbb{E}[\|\nabla c(W)\|^2]}\, \mathcal{W}_2(Z,W) + \frac{M}{2}\, \mathcal{W}_2(Z,W)^2. \tag{25}$$

We bound each term separately. For the quadratic term, the triangle inequality for $\mathcal{W}_2$ gives $\mathcal{W}_2(Z,W) \leq \sigma_Z + \sigma_W$ where $\sigma_Z = \sqrt{\mathbb{E}[\|Z\|^2]}$, so that

$$\frac{M}{2}\mathcal{W}_2(Z,W)^2 \leq \frac{M}{2}(\sigma_Z + \sigma_W)\, \mathcal{W}_2(Z,W). \tag{26}$$

Since h is L-Lipschitz, $\|h(y)\| \leq \|h(0)\| + L\|y\| \leq B_{\mathcal{H}} + L\|y\|$, and therefore

$$\sigma_W^2 = \mathbb{E}[\|h(Y)\|^2 + \|Y\|^2] \leq 2B_{\mathcal{H}}^2 + (1+2L^2){\sigma'}^2 \leq \tau^2,$$

and similarly $\sigma_Z^2 \leq 2B_{\mathcal{H}}^2 + (1+2L^2)\sigma^2 \leq \tau^2$. For the gradient term, since ∇c is M-Lipschitz we have $\|\nabla c(w)\| \leq \|\nabla c(0,0)\| + M\|w\|$, giving

$$\mathbb{E}[\|\nabla c(W)\|^2] \leq 2\|\nabla c(0,0)\|^2 + 2M^2\mathbb{E}[\|W\|^2] \leq 2\big(\|\nabla c(0,0)\|^2 + M^2\tau^2\big).$$

Substituting into (25) using (26),

$$\begin{aligned}
\mathbb{E}[c(Z)] - \mathbb{E}[c(W)] &\leq \left[\sqrt{\mathbb{E}[\|\nabla c(W)\|^2]} + \frac{M}{2}(\sigma_Z + \sigma_W)\right]\mathcal{W}_2(Z,W) \\
&\leq \left[\sqrt{2}\big(\|\nabla c(0,0)\| + M\tau\big) + M\tau\right]\mathcal{W}_2(Z,W) \\
&\leq 3\big(\|\nabla c(0,0)\| + M\tau\big)\, \mathcal{W}_2(Z,W).
\end{aligned}$$

Combining with (24) and exchanging the roles of p_{x} and p'_{x} completes the proof. □

G.2 Stability of the Mixed Risk Minimizer

Denote by $\mathcal{A}$, the algorithm that minimizes the mixed empirical risk:

$$\mathcal{A}(\mathbf{S}) := \operatorname{argmin}_{h\in\mathcal{H}} \mathcal{R}_\lambda(h,\mathbf{S}) = h_{\mathbf{S}}.$$

We develop stability bounds utilizing the loss on synthetic data as a source of regularization. The methodology modifies an approach used in the proof of Proposition 6 of (Farghly et al., 2026), using function space convexity to obtain stability bounds in $L^2(p'_{\mathrm{x}})$ (Lemma G.4) followed by a change of measure argument (Lemma G.5).

Lemma G.3. *Suppose that the function $f : \mathcal{Y} \to \mathbb{R}$ is M-smooth, then for any $y \in \mathcal{Y}$,*

$$f(y) - f^* \geq \frac{1}{2M}\|\nabla f(y)\|^2,$$

where $f^* := \inf_{y\in\mathcal{Y}} f(y)$.

Proof. Let $\langle\cdot,\cdot\rangle$ denote the inner product associated with the space $\mathcal{Y}$. From smoothness, it follows that f is differentiable. Setting $z = y - \frac{1}{M}\nabla f(y)$, it further follows from smoothness that,

$$\begin{aligned}
f(z) - f(y) &\leq \langle \nabla f(y), z-y\rangle + \frac{M}{2}\|z-y\|^2 \\
&\leq -\frac{1}{M}\langle \nabla f(y), \nabla f(y)\rangle + \frac{1}{2M}\|\nabla f(y)\|^2 \\
&\leq -\frac{1}{2M}\|\nabla f(y)\|^2.
\end{aligned}$$

Rearranging and using the fact that $f(z) \geq f^*$ leads to the bound in the statement. □

Lemma G.4. *Suppose Assumption 3.2 holds, then for any $i \in \{1, ..., N\}$ it holds that,*

$$\mathbb{E}\Big[\int \|h(x) - h^i(x)\|^2 p'_{\mathrm{x}}(dx)\Big] \leq \frac{8M}{m^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h, \mathbf{S})] + \frac{4\sqrt{2M}D(1-\lambda)}{mN\lambda}\Big(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2}\Big),$$

where $h = \mathcal{A}(\mathbf{S}), h^i = \mathcal{A}(\mathbf{S}^i)$, $\mathbf{S}^i$ is as defined in Lemma G.1.

Proof. Define the measures,

$$\hat{q}(d\mathrm{x}) := \frac{1-\lambda}{N}\sum_{x_j \in \mathbf{S}} \delta_{\mathrm{x}_j}(d\mathrm{x}) + \lambda p'_{\mathrm{x}}(d\mathrm{x}), \qquad \tilde{q}(d\mathrm{x}) := \frac{1-\lambda}{N}\sum_{\mathrm{x}_j \in \mathbf{S}^i} \delta_{\mathrm{x}_j}(d\mathrm{x}) + \lambda p'_{\mathrm{x}}(d\mathrm{x}).$$

Using the strong convexity of c we obtain,

$$\langle h^i(\mathrm{x}) - h(\mathrm{x}), \nabla_1 c(h^i(\mathrm{x}), \mathrm{x})\rangle \geq c(h^i(\mathrm{x}), \mathrm{x}) - c(h(\mathrm{x}), \mathrm{x}) + \frac{m}{2}\|h^i(\mathrm{x}) - h(\mathrm{x})\|^2.$$

which when integrated with respect to $\hat{q}$, leads to

$$\int \langle h^i(\mathrm{x}) - h(\mathrm{x}), \nabla_1 c(h^i(\mathrm{x}), \mathrm{x})\rangle\, \hat{q}(d\mathrm{x}) \geq \mathcal{R}_\lambda(h^i, \mathbf{S}) - \mathcal{R}_\lambda(h, \mathbf{S}) + \frac{m}{2}\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2 \hat{q}(d\mathrm{x}).$$

The right-hand side is lower bounded further using $\mathcal{R}_\lambda(h^i, \mathbf{S}) \geq \mathcal{R}_\lambda(h, \mathbf{S})$ and the left-hand side is upper bounded using,

$$\begin{aligned}
&\int \langle h^i(\mathrm{x}) - h(\mathrm{x}), \nabla_1 c(h^i(\mathrm{x}), \mathrm{x})\rangle\, \hat{q}(d\mathrm{x}) \\
&\quad = \int \langle h^i(\mathrm{x}) - h(\mathrm{x}), \nabla_1 c(h^i(\mathrm{x}), \mathrm{x})\rangle\, \tilde{q}(d\mathrm{x}) \\
&\qquad + \frac{1-\lambda}{N}\Big(\langle h^i(\mathrm{x}_i) - h(\mathrm{x}_i), \nabla_1 c(h^i(\mathrm{x}_i), \mathrm{x}_i)\rangle - \langle h^i(\tilde{\mathrm{x}}) - h(\tilde{\mathrm{x}}), \nabla_1 c(h^i(\tilde{\mathrm{x}}), \tilde{\mathrm{x}})\rangle\Big) \\
&\quad \leq \Big(\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2\, \tilde{q}(d\mathrm{x})\Big)^{1/2}\Big(\int \|\nabla_1 c(h^i(\mathrm{x}), \mathrm{x})\|^2\, \tilde{q}(d\mathrm{x})\Big)^{1/2} \\
&\qquad + \frac{D(1-\lambda)}{N}\Big(\|\nabla_1 c(h^i(\mathrm{x}_i), \mathrm{x}_i)\| + \|\nabla_1 c(h^i(\tilde{\mathrm{x}}), \tilde{\mathrm{x}})\|\Big) \\
&\quad \leq \sqrt{2M}\Big(\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2\, \tilde{q}(d\mathrm{x})\Big)^{1/2}\mathcal{R}_\lambda(h^i, \mathbf{S}^i)^{1/2} + \frac{\sqrt{2M}D(1-\lambda)}{N}\Big(c(h^i(\mathrm{x}_i), \mathrm{x}_i)^{1/2} + c(h^i(\tilde{\mathrm{x}}), \tilde{\mathrm{x}})^{1/2}\Big).
\end{aligned}$$

The first inequality above follows from the Cauchy–Schwarz inequality and the second follows from Lemma G.3. This results in the bound,

$$\begin{aligned}
\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2 \hat{q}(d\mathrm{x}) \leq{}& \frac{2\sqrt{2M}}{m}\Big(\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2\, \tilde{q}(d\mathrm{x})\Big)^{1/2}\mathcal{R}_\lambda(h^i, \mathbf{S}^i)^{1/2} \\
&+ \frac{2\sqrt{2M}D(1-\lambda)}{mN}\Big(c(h^i(\mathrm{x}_i), \mathrm{x}_i)^{1/2} + c(h^i(\tilde{\mathrm{x}}), \tilde{\mathrm{x}})^{1/2}\Big).
\end{aligned}$$

Taking the expectation, we use the fact that $(h, h^i, \mathbf{S}^i)$ shares the same law as $(h^i, h, \mathbf{S})$ and thus can be exchanged, as well as the symmetry of the algorithm $\mathcal{A}$ under permutations in the dataset, to obtain,

$$\begin{aligned}
\mathbb{E}\Big[\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2 \hat{q}(d\mathrm{x})\Big] \leq{}& \frac{2\sqrt{2M}}{m}\Big(\mathbb{E}\Big[\int \|h^i(\mathrm{x}) - h(\mathrm{x})\|^2\, \hat{q}(d\mathrm{x})\Big]\Big)^{1/2}\mathbb{E}[\mathcal{R}_\lambda(h, \mathbf{S})]^{1/2} \\
&+ \frac{2\sqrt{2M}D(1-\lambda)}{mN}\Big(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2}\Big).
\end{aligned}$$

By solving the quadratic equation, we deduce that this implies,

$$\mathbb{E}\Big[\int \|h^i(x)-h(x)\|^2\hat{q}(dx)\Big]^{1/2} \leq \frac{\sqrt{2M}}{m}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})]^{1/2} + \sqrt{\frac{2M}{m^2}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{2\sqrt{2M}D(1-\lambda)}{mN}\Big(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2}+\mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2}\Big)}.$$

This leads to the bound,

$$\begin{aligned}\mathbb{E}\Big[\int \|h^i(x)-h(x)\|^2 p'_{\mathrm{x}}(dx)\Big] &\leq \frac{1}{\lambda}\mathbb{E}\Big[\int \|h^i(x)-h(x)\|^2\hat{q}(dx)\Big] \\ &\leq \frac{8M}{m^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{4\sqrt{2M}D(1-\lambda)}{mN\lambda}\Big(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2}+\mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2}\Big).\end{aligned}$$

□

Lemma G.5. *Suppose that $\mathcal{H}$ consists of L-Lipschitz functions, then for any $c>1$ and $\delta>0$ sufficiently small, the assumption in Equation 22 is satisfied with*

$$\varepsilon \leq \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + 2MC^{-1}\delta^{-d^*}\sup_i \mathbb{E}\Big[\int \|h(x)-h^i(x)\|^2 p_{\mathrm{x}}(dx)\Big] + 4ML^2\delta^2.$$

Proof. Define $\varepsilon = \mathbb{E}[c(h^i(\mathrm{x}_i),\mathrm{x}_i) - c(h(\mathrm{x}_i),\mathrm{x}_i)]$, then because of the smoothness of c, we obtain

$$\begin{aligned}\varepsilon &\leq \mathbb{E}[\langle h^i(\mathrm{x}_i)-h(\mathrm{x}_i), \nabla_1 c(h(\mathrm{x}_i),\mathrm{x}_i)\rangle] + \frac{M}{2}\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2] \\ &\leq \sqrt{2M}\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2]^{1/2}\mathbb{E}[c(h(\mathrm{x}_i),\mathrm{x}_i)]^{1/2} + \frac{M}{2}\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2] \\ &\leq \sqrt{2M}\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2]^{1/2}\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \frac{M}{2}\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2] \\ &\leq \mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)] + M\mathbb{E}[\|h^i(\mathrm{x}_i)-h(\mathrm{x}_i)\|^2].\end{aligned}$$

Define the measure,

$$p_{\mathrm{x}}^{\tilde{\mathrm{x}},\delta}(dx) := \nVdash_{B_\delta(\tilde{\mathrm{x}})}(x)\, p_{\mathrm{x}}(B_\delta(\tilde{\mathrm{x}}))^{-1}\, p_{\mathrm{x}}(dx).$$

Then we can relate function evaluations to the integral over p'_{x} as follows:

$$\begin{aligned}\|h(\tilde{\mathrm{x}})-h^i(\tilde{\mathrm{x}})\| &\leq \Big(\int \|h(x)-h^i(x)\|^2\, p_{\mathrm{x}}^{\tilde{\mathrm{x}},\delta}(dx)\Big)^{1/2} + \Big(\int \|h(x)-h(\tilde{\mathrm{x}})\|^2\, p_{\mathrm{x}}^{\tilde{\mathrm{x}},\delta}(dx)\Big)^{1/2} \\ &\quad + \Big(\int \|h^i(x)-h^i(\tilde{\mathrm{x}})\|^2\, p_{\mathrm{x}}^{\tilde{\mathrm{x}},\delta}(dx)\Big)^{1/2} \\ &\leq p_{\mathrm{x}}(B_\delta(\tilde{\mathrm{x}}))^{-1/2}\Big(\int \|h(x)-h^i(x)\|^2\, p_{\mathrm{x}}(dx)\Big)^{1/2} + 2L\delta.\end{aligned}$$

Taking the expectation gives,

$$\mathbb{E}[\|h(\mathrm{x}_i)-h^i(\mathrm{x}_i)\|^2]^{1/2} \leq \mathbb{E}_{\tilde{\mathrm{x}}\sim p_{\mathrm{x}}}\Big[p_{\mathrm{x}}(B_\delta(\tilde{\mathrm{x}}))^{-1}\Big]^{1/2}\mathbb{E}\Big[\int \|h(x)-h^i(x)\|^2\, p_{\mathrm{x}}(dx)\Big]^{1/2} + 2L\delta.$$

Using Assumption 3.3, we obtain that for δ sufficiently small,

$$\mathbb{E}[\|h(\mathrm{x}_i)-h^i(\mathrm{x}_i)\|^2]^{1/2} \leq C^{-1/2}\delta^{-d^*/2}\mathbb{E}\Big[\int \|h(x)-h^i(x)\|^2\, p_{\mathrm{x}}(dx)\Big]^{1/2} + 2L\delta.$$

□

G.3 Proof of Theorem 3.1

Theorem 3.1. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, suppose that* $h_{\mathbf{S}} \in \arg\min_{h\in\mathcal{H}} \mathcal{R}_\lambda(h,\mathbf{S})$ *and let* $\mathcal{R}^\star = \min_{h\in\mathcal{H}} r(h)$. *Then, there exists* $N_0, \mathcal{R}_0^\star, \mathcal{W}_0 > 0$ *such that whenever* $N \geq N_0, \mathcal{R}^\star \leq \mathcal{R}_0^\star$ *and* $\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}) \leq \mathcal{W}_0$, *the algorithm* $\mathcal{A}(\mathbf{S}) = h_{\mathbf{S}}$ *is uniformly stable with stability constant,*

$$\varepsilon \lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(h_{\mathbf{S}})] + L^{\frac{2d^*}{d^*+2}} \Delta^{\frac{2}{d^*+2}}, \qquad \Delta = \frac{M}{Cm^2\lambda}\mathcal{R}^\star + \frac{\sqrt{M}D}{Cm\lambda N} + \frac{M\xi}{Cm^2}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}). \tag{27}$$

Furthermore, we have the generalization bound,

$$\mathbb{E}[r(h_{\mathbf{S}})] \lesssim \mathcal{R}^\star + \lambda\xi\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)L^{\frac{2d^*}{d^*+2}}\Delta^{\frac{2}{d^*+2}}.$$

Proof. To obtain the stability bound in the first part of the theorem, we combine lemmas G.4 and G.5. We optimize δ in the bound of Lemma G.5, such that

$$\delta^{d^*+2} = \frac{d^*K}{2L^2}, \qquad K = C^{-1}\sup_i \mathbb{E}\left[\int \|h(x) - h^i(x)\|^2 p_{\mathrm{x}}(dx)\right],$$

so that for K sufficiently small, we obtain stability with constant,

$$\varepsilon \lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + K^{\frac{2}{d^*+2}} L^{\frac{2d^*}{d^*+2}}. \tag{28}$$

To control K we use that h and h^i are Lipschitz and thus,

$$\begin{aligned} K &\leq C^{-1}\sup_i \left\{\mathbb{E}\left[\int \|h(\tilde{x}) - h^i(\tilde{x})\|^2 p'_{\mathrm{x}}(d\tilde{x})\right]^{\frac{1}{2}} + 2\mathbb{E}\left[\int \|h^i(x) - h^i(\tilde{x})\|^2 \pi(dx, d\tilde{x})\right]^{\frac{1}{2}}\right\}^2 \\ &\leq 2C^{-1}\sup_i \mathbb{E}\left[\int \|h(x) - h^i(x)\|^2 p'_{\mathrm{x}}(dx)\right] + 8C^{-1}L^2\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2 \end{aligned}$$

which when substituting the bound from Lemma G.4, produces,

$$K \leq \frac{16M}{Cm^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{8\sqrt{2M}D(1-\lambda)}{Cm\lambda N}\left(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2}\right) + \frac{8L^2}{C}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2.$$

Since ℓ is smooth and $\mathcal{H}$ has bounded diameter, we have that $\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]$ and $\mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]$ are uniformly bounded over $h \in \mathcal{H}$. Therefore, it follows that K can be taken sufficiently small as soon as N^{-1}, $\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})]$ and $\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})$ are all taken sufficiently small. From Lemma G.1, it follows that $\mathbb{E}[\mathcal{L}_{\mathcal{X}}(h)]^{1/2} \leq \mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \varepsilon^{1/2}$. Therefore, substituting back into (28), we obtain,

$$\begin{aligned} \varepsilon &\lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + \left(\frac{ML^{d^*}}{Cm^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{\sqrt{M}DL^{d^*}(1-\lambda)}{Cm\lambda N}\left(\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \varepsilon^{1/2}\right) + \frac{L^{2+d^*}}{C}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2\right)^{\frac{2}{d^*+2}} \\ &\lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + \left(\frac{ML^{d^*}}{Cm^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{\sqrt{M}DL^{d^*}(1-\lambda)}{Cm\lambda N}\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \frac{L^{2+d^*}}{C}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2\right)^{\frac{2}{d^*+2}} \\ &\quad + \left(\frac{\sqrt{M}DL^{d^*}(1-\lambda)}{Cm\lambda N}\right)^{\frac{2}{d^*+2}}\varepsilon^{\frac{1}{d^*+2}}. \end{aligned}$$

It then follows from an application of Young's inequality,

$$\begin{aligned} \varepsilon &\lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + \left(\frac{ML^{d^*}}{Cm^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{\sqrt{M}DL^{d^*}(1-\lambda)}{Cm\lambda N}\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)]^{1/2} + \frac{L^{2+d^*}}{C}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2\right)^{\frac{2}{d^*+2}} \\ &\quad + \left(\frac{\sqrt{M}DL^{d^*}(1-\lambda)}{Cm\lambda N}\right)^{\frac{2}{d^*+1}} \\ &\lesssim \mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + \left(\frac{ML^{d^*}}{Cm^2\lambda}\mathbb{E}[\mathcal{R}_\lambda(h,\mathbf{S})] + \frac{\sqrt{M}DL^{d^*}}{Cm\lambda N} + \frac{L^{2+d^*}}{C}\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}})^2\right)^{\frac{2}{d^*+2}}, \end{aligned}$$

where in the last line, we use $N \geq \frac{\sqrt{M}DL^{d^*}}{Cm\lambda}$ and $(1-\lambda)\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)] \leq \mathbb{E}[\mathcal{R}_\lambda(h, \mathbf{S})] \leq 1$, by assumption.

To simplify the bound, we also control $\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S})$. For any $h_\star \in \mathcal{H}$, by the optimality of $h_{\mathbf{S}}$, we have

$$\begin{aligned}\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S}) &\leq \mathcal{R}_\lambda(h_\star, \mathbf{S}),\\ \mathbb{E}_{\mathbf{S}}[\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S})] &\leq \mathbb{E}_{\mathbf{S}}[\mathcal{R}_\lambda(h_\star, \mathbf{S})],\\ \mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S}) &\leq r_\lambda(h_\star).\end{aligned}$$

From the definition of $r_\lambda(h_\star)$ (see Equation 3), we can write:

$$\begin{aligned}\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S}) &\leq (1-\lambda)\mathbb{E}_{p_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})] + \lambda\mathbb{E}_{p'_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})]\\ &= r(h_\star) + \lambda\left(\mathbb{E}_{p'_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})] - \mathbb{E}_{p_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})]\right).\end{aligned} \tag{29}$$

Since $\ell(h, \mathbf{x})$ is ξ-Lipschitz in $\mathbf{x}$, it follows from Lemma G.2 that,

$$\mathbb{E}_{p'_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})] - \mathbb{E}_{p_{\mathbf{x}}}[\ell(h_\star, \mathbf{x})] \leq \xi\mathcal{W}_2\left(p_{\mathbf{x}}\ , p'_{\mathbf{x}}\right),$$

and thus,

$$\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S}) \leq r(h_\star) + \xi\lambda\mathcal{W}_2\left(p_{\mathbf{x}}\ , p'_{\mathbf{x}}\right).$$

Finally, we can choose $h_\star = \arg\min_{h\in\mathcal{H}} r(h)$ to tighten the bound. With this, we conclude the proof of the bound in (27), which holds once $N^{-1}, \mathcal{R}^\star, \mathcal{W}_2\left(p_{\mathrm{x}}\ , p'_{\mathrm{x}}\right)$ are all taken sufficiently small.

For the second bound of the theorem, we use the following decomposition to upper bound the generalization error:

$$r(h) = [r(h) - r_\lambda(h)] + [r_\lambda(h) - \mathcal{R}_\lambda(h, \mathbf{S})] + \mathcal{R}_\lambda(h, \mathbf{S}).$$

We proceed by bounding $r(h) - r_\lambda(h)$. We compute

$$\begin{aligned}r(h) - r_\lambda(h) &= \mathbb{E}_{p_{\mathbf{x}}}[\ell(h, \mathbf{x})] - (1-\lambda)\mathbb{E}_{p_{\mathbf{x}}}[\ell(h, \mathbf{x})] - \lambda\mathbb{E}_{p'_{\mathbf{x}}}[\ell(h, \mathbf{x})]\\ &= \lambda\left(\mathbb{E}_{p_{\mathbf{x}}}[\ell(h, \mathbf{x})] - \mathbb{E}_{p'_{\mathbf{x}}}[\ell(h, \mathbf{x})]\right)\\ &\leq \lambda\xi\,\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}),\end{aligned}$$

where the inequality follows from Lemma G.2.

To bound $r_\lambda(h) - \mathcal{R}_\lambda(h, \mathbf{S})$, we start by substituting the definition of each term and simplifying them. We have:

$$\begin{aligned}r_\lambda(h) - \mathcal{R}_\lambda(h, \mathbf{S})) &= (1-\lambda)\left(\mathbb{E}_{\mathbf{S},\mathbf{x}'}\left[\ell(h_{\mathbf{S}}, \mathbf{x}')\right] - \mathbb{E}_{\mathbf{S},i}\left[\ell(h_{\mathbf{S}}, \mathbf{x}_i)\right]\right)\\ &= (1-\lambda)\left(\mathbb{E}_{\mathbf{S},\mathbf{x}'}\left[\ell(h_{\mathbf{S}'}, \mathbf{x}_i)\right] - \mathbb{E}_{\mathbf{S},i}\left[\ell(h_{\mathbf{S}}, \mathbf{x}_i)\right]\right),\end{aligned}$$

where the first equality is due to the fact that $\lambda\mathbb{E}_{p'_{\mathbf{x}}}[\ell(h, \mathbf{x})]$ is common in both terms. The second equality is by the definition of each term, and the fact that $\mathbf{x}' \perp\!\!\!\perp \mathbf{S}$. Note that $i \sim \mathrm{Unif}([N])$. The final line results from defining $\mathbf{S}' = \mathbf{S} \cup \{\mathbf{x}'\} \setminus \{\mathbf{x}_i\}$, which is a neighboring set to $\mathbf{S}$. Now, assuming that we have ε-uniformly stable algorithm $\mathcal{A}$, then we can write

$$\begin{aligned}r_\lambda(h) - \mathcal{R}_\lambda(h, \mathbf{S}) &= (1-\lambda)\mathbb{E}_{\mathbf{S},\mathbf{x}',i}\left[\ell(h_{\mathbf{S}'}, \mathbf{x}_i) - \ell(h_{\mathbf{S}}, \mathbf{x}_i)\right]\\ &\leq (1-\lambda)\varepsilon\,.\end{aligned}$$

These combine to give the bound,

$$\mathbb{E}[r(h) - r(h_\star)] \leq 2\lambda\xi\,\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)\varepsilon.$$

Substituting (27) into the bound above, we obtain,

$$\mathbb{E}[r(h_{\mathbf{S}}) - r(h_\star)] \lesssim 2\lambda\xi\,\mathcal{W}_2(p_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)\mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] + (1-\lambda)L^2\Delta^{\frac{2}{d^*+2}}.$$

Since $(1-\lambda)\mathbb{E}[\mathcal{L}_{\mathbf{S}}(h)] \leq \mathbb{E}[\mathcal{R}_\lambda(h, \mathbf{S})]$, and using $\mathbb{E}[\mathcal{R}_\lambda(h_{\mathbf{S}}, \mathbf{S})] \leq \mathcal{R}^\star + \xi\lambda\mathcal{W}_2\left(p_{\mathbf{x}}\ , p'_{\mathbf{x}}\right)$ from above, we obtain,

$$(1-\lambda)\mathbb{E}[\mathcal{L}_{\mathbf{S}}(\mathcal{A}(\mathbf{S}))] \leq \mathcal{R}^\star + \xi\lambda\mathcal{W}_2\left(p_{\mathbf{x}}\ , p'_{\mathbf{x}}\right).$$

□

H THEORETICAL RESULTS AND DISCUSSIONS OF DOMAIN SHIFT

H.1 Proof of Theorem 5.1

Theorem 5.1. *Under Assumption 2.1, for the kernel regression problem defined in Equation 1 and any fixed regularization parameter $\lambda > 0$, the test error under domain shift satisfies the bound:*

$$\mathcal{R}_N(\lambda; g) = \mathcal{O}\left(\lambda^{-2}\mathcal{D}(f_\star, \tilde{f}) + \lambda^{2-\frac{1}{4r}}\mathcal{D}(f_\star, g) + \frac{\sigma^2 + \mathcal{D}(f_\star, \tilde{f}) + \mathcal{D}(f_\star, g)}{N\lambda^2}\right),$$

where $\mathcal{D}(\cdot,\cdot)$ denotes the distributional discrepancy, as in Theorem 2.2.

Proof. The proof follows the proof of Theorem 2.2 in Appendix F.2, using the bias-variance decomposition. We note that the variance term remains the same as it only depends on the noise of the data, while the bias term will have the dependency on all three terms of $f_\star$, $\tilde{f}$ and g. Let us start with the formal definition of our bias term, similar to Appendix F.2:

$$\begin{aligned}
\mathcal{B}^2 &= \mathbb{E}_{\mathbf{x}}\left[(f_\star - \mathbb{E}_\varepsilon[f_N])(\mathrm{x})\right]^2 \\
&\leq \mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon[f_N] - f_\lambda\|^2_{\mathcal{H}_K}\right] + \|f_\star - f_\lambda\|^2_{p_{\mathbf{x}}} \\
&\leq \mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon[f_N] - \tilde{f}_\lambda\|^2_{\mathcal{H}_K}\right] + \|\tilde{f}_\lambda - f_\lambda\|^2_{p_{\mathbf{x}}} + \|f_\star - f_\lambda\|^2_{p_{\mathbf{x}}},
\end{aligned}$$

where $\tilde{f}_\lambda$, and f_λ are the population limits of f_N with repsect to $\tilde{f}$ and $f_\star$, i.e. $\tilde{f}_\lambda = g + (T_K + \lambda I)^{-1}T_K(\tilde{f} - g)$, and $f_\lambda = g + (T_K + \lambda I)^{-1}T_K(f_\star - g)$. We have previously studied the first term in Equation 20. Therefore, we have:

$$\mathbb{E}_{\mathbf{S}}\left[\|\mathbb{E}_\varepsilon[f_N] - f_\lambda\|^2_{\mathcal{H}_K}\right] \leq \frac{3\kappa^4\|\tilde{f} - g\|^2_{\mathcal{H}}}{N\lambda^2} \leq \frac{3\kappa^4}{N\lambda^2}\left(\|f_\star - \tilde{f}\|^2_{\mathcal{H}} + \|\tilde{f} - g\|^2_{\mathcal{H}}\right) \leq \frac{3\kappa^4}{N\lambda^2}\left(\mathcal{D}(f_\star, \tilde{f}) + \mathcal{D}(f_\star, g)\right).$$

The last term follows from Equation 21, and the computation after that. Therefore, we have:

$$\begin{aligned}
\|f_\star - f_\lambda\|^2_{p_{\mathbf{x}}} &\leq \kappa^2\|f_\star - f_\lambda\|^2_{\mathcal{H}} \\
&\leq \kappa^2\sum_{j=1}^{\infty}\frac{\lambda^2}{(\mu_j + \lambda)^2}(\theta_j - \omega_j)^2 \\
&\leq \kappa^2\lambda^2\sqrt{\left(\sum_{j=1}^{\infty}\frac{\mu_j^2}{(\mu_j + \lambda)^2}\right)\left(\sum_{j=1}^{\infty}\frac{1}{\mu_j^2}(\theta_j - \omega_j)^2\right)} \\
&= \mathcal{O}\left(\kappa^2\lambda^{2-\frac{1}{4r}}\mathcal{D}(f_\star, g)C_r\right).
\end{aligned}$$

Now, it's only enough to provide a bound for the second term. We have:

$$\begin{aligned}
\|\tilde{f}_\lambda - f_\lambda\|^2_{p_{\mathbf{x}}} &\leq \kappa^2\|\tilde{f}_\lambda - f_\lambda\|^2_{\mathcal{H}} \\
&= \kappa^2\|(T_K + \lambda I)^{-1}T_K(f_\star - \tilde{f})\|^2_{\mathcal{H}} \\
&\leq \kappa^2\frac{\kappa^2}{\lambda^2}\|f_\star - \tilde{f}\|^2_{\mathcal{H}} \\
&\leq \frac{\kappa^4}{\lambda^2}\mathcal{D}(f_\star, \tilde{f}).
\end{aligned}$$

Combining all the terms completes the proof. □

H.2 Proof of Theorem 5.2

Theorem 5.2. *Suppose that Assumptions 3.1–3.3 hold and that $h_{\mathbf{S}} \in \arg\min_{h\in\mathcal{H}}\mathcal{R}_\lambda(h, \mathbf{S})$. Then, whenever $N \geq N_0, \mathcal{R}^\star \leq \mathcal{R}_0^\star$ and $\mathcal{W}_2(p_{\mathrm{x}}, p'_{\mathrm{x}}) \leq \mathcal{W}_0$, the generalization gap under the domain shift satisfies*

$$\mathbb{E}[r(h_{\mathbf{S}})] \lesssim r^\star + \lambda\xi\mathcal{W}_2(p^\star_{\mathbf{x}}, p'_{\mathbf{x}}) + (1-\lambda)\xi\mathcal{W}_2(p^\star_{\mathbf{x}}, p_{\mathbf{x}}) + (1-\lambda)L^{\frac{2d^*}{d^*+2}}\Delta^{\frac{2}{d^*+2}},$$

where $r^\star = \min_{h\in\mathcal{H}} r^\star(h)$ is minimum population risk minimizer of the target domain and $\Delta, \xi, N_0, \mathcal{R}_0^\star, \mathcal{W}_0$ are as in Theorem 3.1.

Proof. For any $h \in \mathcal{H}$, we show $r^\star(h)$, and $r_\lambda^\star(h)$ as

$$r^\star(h) = \mathbb{E}_{p_{\mathbf{x}}^\star}\left[\ell(h,\mathbf{x})\right], \qquad r_\lambda^\star(h) = (1-\lambda)r^\star(h) + \mathbb{E}_{p'_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]. \tag{30}$$

We now have the following decomposition for the generalization error:

$$r^\star(h) = (r^\star(h) - r_\lambda^\star(h)) + (r_\lambda^\star(h) - r_\lambda(h)) + (r_\lambda(h) - \mathcal{R}_\lambda(h)) + \mathcal{R}_\lambda(h).$$

Similar to Appendix G.3, we bound each of the terms separately. We have:

Bounding $r^\star(h) - r_\lambda^\star(h)$: Let us expand the term by the definition of each component:

$$\begin{aligned} r^\star(h) - r_\lambda^\star(h) &= \mathbb{E}_{p_{\mathbf{x}}^\star}\left[\ell(h,\mathbf{x})\right] - \left((1-\lambda)r^\star(h) + \mathbb{E}_{p'_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]\right) \\ &= \lambda\left(\mathbb{E}_{p_{\mathbf{x}}^\star}\left[\ell(h,\mathbf{x})\right] - \mathbb{E}_{p'_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]\right) \\ &\le \lambda\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p'_{\mathbf{x}}\right), \end{aligned}$$

where the last inequality is by Lemma G.2.

Bounding $r_\lambda^\star(h) - r_\lambda(h)$: We again use the definitions:

$$\begin{aligned} r_\lambda^\star(h) - r_\lambda(h) &= \left((1-\lambda)r^\star(h) + \mathbb{E}_{p'_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]\right) - \left((1-\lambda)r(h) + \mathbb{E}_{p'_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]\right) \\ &= (1-\lambda)(r^\star(h) - r(h)) \\ &= (1-\lambda)\left(\mathbb{E}_{p_{\mathbf{x}}^\star}\left[\ell(h,\mathbf{x})\right] - \mathbb{E}_{p_{\mathbf{x}}}\left[\ell(h,\mathbf{x})\right]\right) \\ &\le (1-\lambda)\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p_{\mathbf{x}}\right), \end{aligned} \tag{31}$$

where we have once again used Lemma G.2.

Bounding $r_\lambda(h) - \mathcal{R}_\lambda(h)$: Similar to Appendix G.3, we refer to this term as the stability term. Note that all the conditions for Lemma G.1 hold here, therefore, this term is the same as Appendix G.3 since the stability is uniform. Thus, $r_\lambda(h) - \mathcal{R}_\lambda(h) \le (1-\lambda)\varepsilon$ for:

$$\varepsilon \lesssim \frac{1}{1-\lambda}\mathbb{E}[\mathcal{R}_\lambda(h_{\mathbf{S}})] + L^{\frac{2d^*}{d^*+2}}\Delta^{\frac{2}{d^*+2}}. \tag{32}$$

We now only need to bound $\mathcal{R}_\lambda(h_{\mathbf{S}})$, for both Equations 30 and 32.

Bounding $\mathcal{R}_\lambda(h_{\mathbf{S}})$: Since $h_{\mathbf{S}} = \arg\min_{h\in\mathcal{H}} \mathcal{R}_\lambda(h,\mathbf{S})$ is the empirical minimizer, for any $h' \in \mathcal{H}$, by optimality of $h_{\mathbf{S}}$, we have

$$\begin{aligned} \mathcal{R}_\lambda(h_{\mathbf{S}},\mathbf{S}) &\le \mathcal{R}_\lambda(h',\mathbf{S}), \\ \mathbb{E}_{\mathbf{S}}[\mathcal{R}_\lambda(h_{\mathbf{S}},\mathbf{S})] &\le \mathbb{E}_{\mathbf{S}}[\mathcal{R}_\lambda(h',\mathbf{S})], \\ \mathcal{R}_\lambda(h_{\mathbf{S}},\mathbf{S}) &\le r_\lambda(h'). \end{aligned}$$

Now, let $h' = \arg\min_{h\in\mathcal{H}} r_\lambda(h)$. Then, for any $h_\star \in \mathcal{H}$, we have

$$\begin{aligned} \mathcal{R}_\lambda(h_{\mathbf{S}},\mathbf{S}) &\le r_\lambda(h') \\ &\le r_\lambda^\star(h_\star) + (1-\lambda)\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p_{\mathbf{x}}\right) \\ &\le r^\star(h_\star) + \lambda\left(\mathbb{E}_{p_{\mathbf{x}}^\star}[\ell(h_\star,\mathbf{x})] - \mathbb{E}_{p'_{\mathbf{x}}}[\ell(h_\star,\mathbf{x})]\right) + (1-\lambda)\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p_{\mathbf{x}}\right) \\ &\le r^\star(h_\star) + \lambda\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p'_{\mathbf{x}}\right) + (1-\lambda)\xi\mathcal{W}_2\left(p_{\mathbf{x}}^\star, p_{\mathbf{x}}\right), \end{aligned}$$

where the first inequality is by Equation 31, and the second inequality is from the definition and the last one is by Lemma G.2. Now, let $h_\star = \arg\min_{h\in\mathcal{H}} r^\star(h)$. Combining all bounds together completes the proof. □

I EXPERIMENTAL SETUP

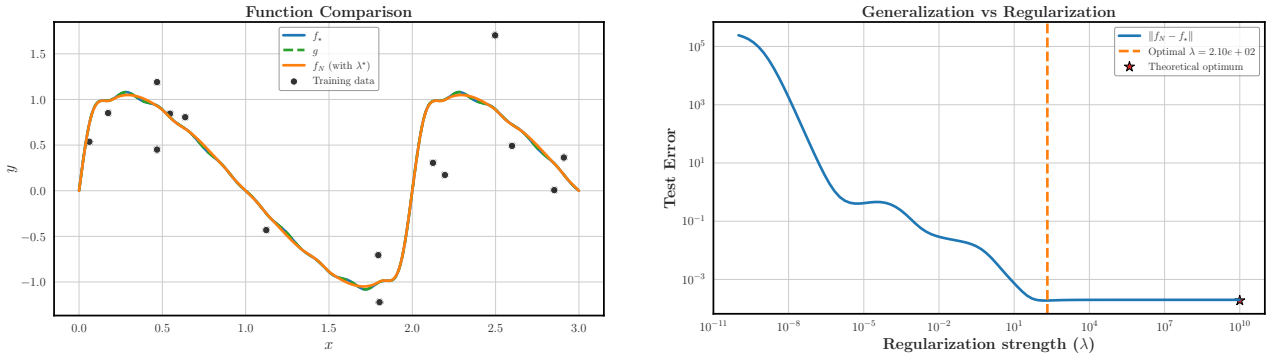
I.1 Optimal Regularization in Kernel Ridge Regression

We study a nonparametric regression problem wherein the ground truth function $f_\star$ and an auxiliary function g are both defined as truncated series expansions in an orthonormal sine basis, with polynomially decaying coefficients to encode varying degrees of smoothness. The target function is given by $f_\star(x) = \sum_{j=1}^{T_f}(j+1)^{-rs}\sin(\pi(j+1)x)$, while g is constructed analogously using a decay rate s' over the first T_g terms. Training data consists of $N = 15$ i.i.d. samples $\{x_i\}_{i=1}^n$ drawn uniformly from $[0, 3]$, with noisy observations $y_i = f_\star(x_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, 0.1)$, alongside noiseless evaluations of $g(x_i)$. We employ our modified kernel ridge regression (Lemma 2.1) method using a Mercer kernel with eigenvalue decay $\mu_j \asymp j^{-2r}$, incorporating g as a regularization term to enhances the standard kernel estimator. The predictive performance is evaluated on a dense test grid (test set of 500 points) by computing the empirical L_2-distance between the learned function f_N and the true function $f_\star$. This procedure is repeated across a logarithmically spaced range of regularization parameters $\lambda \in [10^{-10}, 10^{10}]$. In addition, we compute the theoretically optimal regularization parameter by minimizing an upper bound derived from the distance $\mathcal{D}(f_\star, g)$, which depends explicitly on the eigendecay and coefficient mismatch between $f_\star$ and g, based on Theorem 2.2.

Our implementation uses SciPy for numerical integration and optimization, with special care given to numerical stability through pseudo-inverses and adaptive regularization. To capture the effect of the difference between $f_\star$ and g, we run the experiment with various values as depicted in Table 1. Figures 1, 5 and 6 illustrate the impact of distributional alignment between the true function $f_\star$ and the synthetic generator g on the behaviour of the estimated function f_N and the choice of regularization strength λ. In Figure 5, the synthetic generator perfectly matches the true distribution ($s = s'$, and $T_f = T_g$), resulting in no discrepancy between $f_\star$, g, and f_N. Consequently, the prediction error $\|f_N - f_\star\|_{L_2}$ is minimized for the largest possible regularization strength, and our algorithm successfully selects this value. In contrast, Figure 6 considers a case with distribution mismatch (large difference between s, s', and T_f, T_g), leading to larger discrepancies between the functions. This results in a characteristic U-shaped prediction error curve, as shown in Figure 6b. While the theoretically chosen regularization strength (star marker) slightly overestimates the empirical optimum (dashed orange line), the difference remains negligible, demonstrating the robustness of our theoretical bound under mismatch. The experimental details are shown in Table 1.

Table 1: Effective parameters for the modified kernel regression.

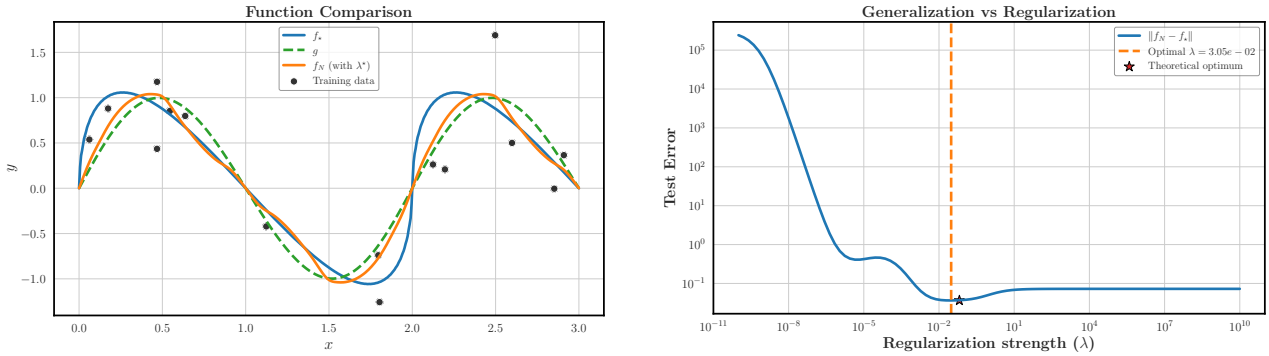
r	s	s'	T_f	T_g	$D(f_\star, g)$	Figure
2.0	0.8	0.8	100	100	0.0000	Figure 5
2.0	0.8	1.5	100	10	737.65	Figure 1
2.0	0.8	2.5	100	10	15509.16	Figure 6



(a) Function comparisons.

(b) Generalization vs Regularization.

Figure 5: *(a)* Comparison of the true function $f_\star$ (blue), the synthetic generator g (green), and the estimated function f_N (orange), obtained via Lemma 2.1, with parameters $r = 2.0$, $s = 0.8$, and $s' = 0.8$. Since $g = f_\star$ in this setting, the RKHS distance is zero and all curves coincide. *(b)* Prediction error $\|f_N - f_\star\|_{L_2}$ as a function of the regularization strength λ. As expected, there is no U-shaped behaviour since the generator fully matches the true distribution. The theoretical optimum selects a large λ (star marker), while the empirical optimum (dashed orange line) selects a smaller value due to numerical precision limits.



(a) Function comparisons.

(b) Generalization vs Regularization.

Figure 6: *(a)* Comparison of the true function $f_\star$ (blue), the synthetic generator g (green), and the estimated function f_N (orange), obtained via Lemma 2.1, with parameters $r = 2.0$, $s = 0.8$, and $s' = 2.5$. The distance between the functions is larger compared to Figure 1. *(b)* Prediction error $\|f_N - f_\star\|_{L_2}$ as a function of the regularization strength λ. A clear U-shaped curve is observed, and while the theoretically optimal λ (star marker) slightly overestimates the empirical optimum (dashed orange line), the difference is negligible.

I.2 Natural Images on CIFAR-10

We investigate the effect of synthetic data on classification performance using a conditional diffusion model. Specifically, we train a diffusion model on CIFAR-10 to generate class-conditional synthetic images, which are then used to augment the real training set. We compare two classifiers: one trained solely on real data, and another trained on a mixture of real and synthetic samples. Performance is evaluated across varying synthetic-to-real data ratios, and validation accuracy is reported for each configuration. The real dataset used to train the diffusion model is disjoint from the one used for training and validating the classifier, allowing us to isolate the effect of synthetic data augmentation. Detailed experimental settings are provided in Appendix I.2.1.

In Figure 7, we observe that classification accuracy improves with increasing amounts of mixed training data when the distributional distance between the synthetic generator and real data is small (orange line in Appendix I.2.1). In contrast, for generators with moderate to high distributional distance (i.e., lower quality - see green and red lines), we observe diminishing returns or even performance degradation. This follows our insights from Section 3. Similar trends are observed at the class level, although the results are noisier due to the reduced amount of data per class—approximately one-tenth of the total. These results indicate that the trained diffusion model captures different classes with varying fidelity, which in turn affects per-class generalization. This highlights an important practical consideration: when class-wise generalization is a priority, it is crucial to ensure that the synthetic data generator performs well not only in aggregate but also across different classes or groups.

I.2.1 Experimental Details for CIFAR-10

Dataset and preprocessing We conduct experiments on CIFAR-10 Krizhevsky et al. (2009), a dataset of 60,000 colour images (32×32 pixels) across 10 object categories, with 50,000 training and 10,000 test samples. For each run, we stratify the training set to construct three disjoint subsets: a labelled training set $\mathcal{D}_{\text{train}}$ containing N examples, a validation set $\mathcal{D}_{\text{val}}$ of 5,000 examples, and a separate set $\mathcal{D}_{\text{diff}}$ of 50,000 examples used for training the diffusion model. Stratified sampling ensures class balance across all subsets. All images are linearly rescaled to the $[-1, 1]$ range. During classifier training, we apply random horizontal flips as the only form of data augmentation. No augmentations are used during diffusion model training or validation.

Conditional diffusion model Our synthetic data generator is a class-conditional diffusion model trained on $\mathcal{D}_{\text{diff}}$. The architecture is a UNet2D with six downsampling and upsampling blocks, using channel sizes [128, 128, 256, 256, 512, 512]. Self-attention layers are included at the 16×16 spatial resolution. Class conditioning is achieved via a learnable embedding table of dimension 512. We train the model using a linear noise schedule over $T = 1000$ diffusion steps. Optimization is performed with AdamW using a learning rate of 10^{-4} and $(\beta_1, \beta_2) = (0.9, 0.999)$. We apply cosine learning rate decay with 500 warmup steps, use mixed-precision training (FP16), and set the batch size to 64. Each model is trained for 100 epochs.

Classification task For the downstream task, we use a compact convolutional neural network. It consists of two convolutional layers with 3×3 kernels and output channels 32 and 64, respectively, each followed by ReLU activation and max pooling. The output is flattened and passed through a fully connected layer with 512 units, followed by ReLU, a dropout layer with rate 0.25, and a final fully connected layer with 10 outputs. We train this classifier using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 64, and up to 20 epochs with early stopping based on validation performance. Cross-entropy loss is used for optimization.

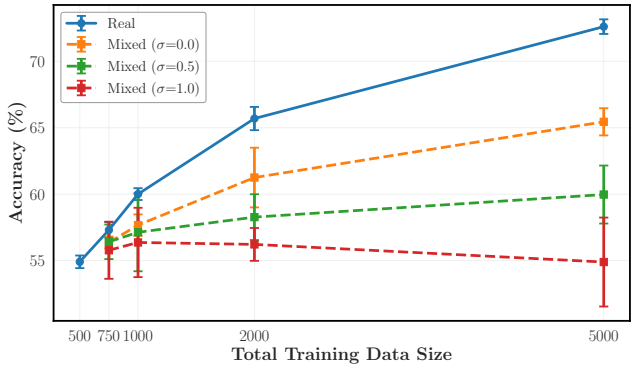
Experimental protocol For each configuration (N, M), where M denotes the number of synthetic samples to generate, we first train the conditional diffusion model on $\mathcal{D}_{\text{diff}}$. We then sample M class-conditional synthetic images to form $\mathcal{D}_{\text{synth}}$. Two classifiers are trained: f_{real} on $\mathcal{D}_{\text{train}}$ alone, and f_{aug} on the augmented dataset $\mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{synth}}$. Both classifiers are evaluated on the same validation set $\mathcal{D}_{\text{val}}$ using the classification accuracy:

$$\text{Acc} = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{(x,y)\in\mathcal{D}_{\text{val}}} \mathbb{I}\left(f(x) = y\right).$$

Hyperparameter configurations We explore several synthetic-to-real data ratios $M/N \in \{0.25, 0.5, 0.75, 1.0, 1.25, 1.5, 2.0, 5.0\}$, with each configuration repeated across multiple random seeds. A summary of the hyperparameters is provided in Table 2. All experiments are implemented with the HuggingFace Diffusers library and executed on NVIDIA A100 GPUs with 40GB memory.

Table 2: Hyperparameter configurations for CIFAR-10 experiments

Parameter	Values
Real data size (N)	500
Synthetic-to-real ratio (M/N)	0.5, 1.0, 3, 9
Diffusion steps (T)	1000
Total noise variances	0.0, 0.5, 1.0



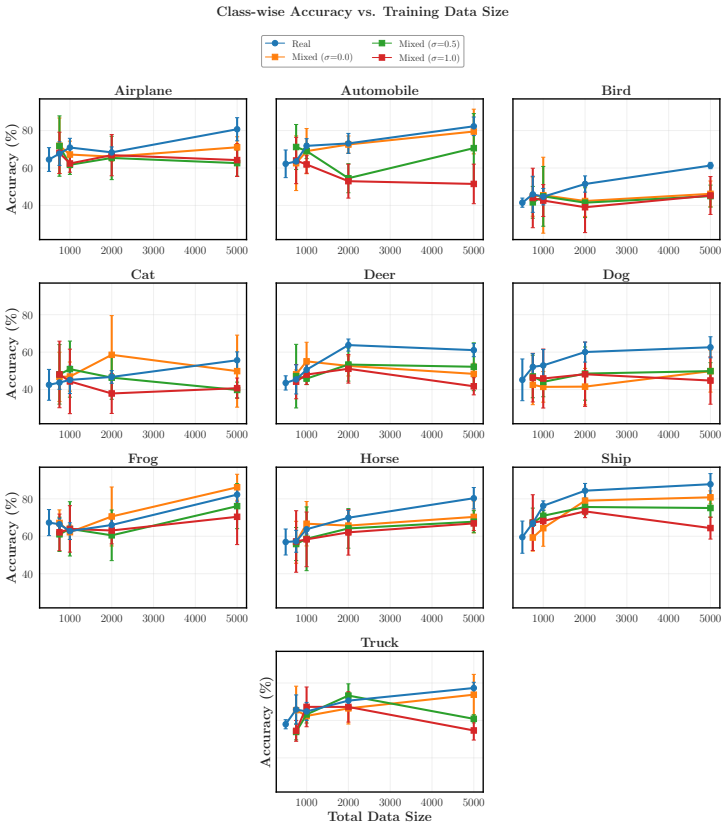


Figure 7: (a) Average accuracy vs. training data size. Increasing the amount of real data (blue line) consistently increases the accuracy of the classification, while for the mixed-data it depends on the quality of the generated samples. (b) Accuracy of each class vs. training data size. We observe a similar pattern, however noisier.

I.3 Real-World Medical Imaging

In this section, we provide additional details for the experimental setup of Section 4.

Diffusion Model Training Our conditional diffusion model synthesises MRI slices conditioned on anatomical tissue masks. The architecture begins with a 1×1 convolutional layer for initial feature projection, followed by a sinusoidal positional embedding to encode timestep information. The model includes four down-sampling stages, each consisting of two ResNet blocks, a linear attention layer with residual connection, and a 3×3 convolutional down-sampling layer. This is followed by a bottleneck module comprising two additional ResNet blocks and another linear attention layer. The up-sampling path mirrors the down-sampling structure, replacing down-sampling layers with convolutional up-sampling layers of the same kernel size. Finally, a 1×1 convolutional layer projects the features to the desired output channels.

The model follows a hierarchical channel structure, starting with 64 channels, doubling at each down-sampling stage ($64 \to 128 \to 256 \to 512$ at the bottleneck), then halving symmetrically during up-sampling back to 64. Conditioning is achieved by concatenating a four-channel binary mask (GM, WM, CSF, lesion) with the timestep embedding and spatial inputs at the input layer. The model is trained on slices from the NeuroRx dataset using mean squared error loss over 600 denoising timesteps. All modalities (T1, T2, PD) are trained independently.

Segmentation Model and Task A vanilla U-Net is used for lesion segmentation. The network comprises three downsampling and upsampling layers with skip connections, ReLU activations, and max pooling. Feature channels double in the downsampling path ($64 \to 128 \to 256$) and halve in the upsampling path symmetrically. For all experiments, the models train for up to 800 epochs or until plateau, validated on a fixed NeuroRx set.

Hyperparameter Configuration We perform a targeted grid search for hyperparameters considering our hardware constraints. The search space and the final configuration is shown in Table 3. All experiments are executed on NVIDIA A100 GPUs with 40GB memory, with no gradient clipping or additional augmentations.

Computational Cost Comparison The full sweep (used to simulate the behaviour of cross-validation) involves training across 8 different data compositions, with total training time across all runs adding up to approximately 222 GPU-hours. In contrast, following our theory requires estimating certain statistical quantities and evaluating the bound once. The cost of these operations is negligible compared to training, and only a single model needs to be trained at the predicted optimal ratio. Hence, our method can save up to 90% of the training cost (1 model instead of 8), while still achieving near-optimal performance. The computational cost (training time) of running our experiments in shown in Table 4.

Table 3: Hyperparameter configurations for medical imaging experiments

Parameter	Values / Search Space
Batch size	16, 32, 64, 128 (selected: 128)
Learning rate	{1e-4, 5e-4, 1e-3} (selected: 1e-4)
Epochs	Up to 800 or until loss plateau
Optimizer	Adam
LR Scheduler	Exponential decay ($\gamma = 0.99$)
Weight Init	Kaiming Uniform
Loss Function	Focal + Tversky loss (equal weight)
Focal Loss Params	$\delta = 0.25$, $\lambda = 2$
Tversky Loss Params	$\alpha = 0.7$, $\beta = 0.3$
Gradient Clipping	None

I.3.1 Additional Results

From Theorem 3.1, we expect that as the quality of synthetic samples deteriorates, i.e., as the distributional distance between the synthetic data generator and the real data increases, the optimal synthetic-to-real ratio should decrease, placing greater emphasis on the real data. Consequently, we anticipate an increase in the

Table 4: Computation time as a function of $N + M$.

$N + M$	**Time (hours)**
4500	∼12
5625	∼15
6750	∼15
9000	∼24
13500	∼36
22500	∼50
40500	∼70

validation loss. Figure 8 empirically supports this expectation. In our setting, this distributional distance is modulated by the sampling timestep of the diffusion model: higher timesteps correspond to noisier, and thus less realistic, synthetic samples.

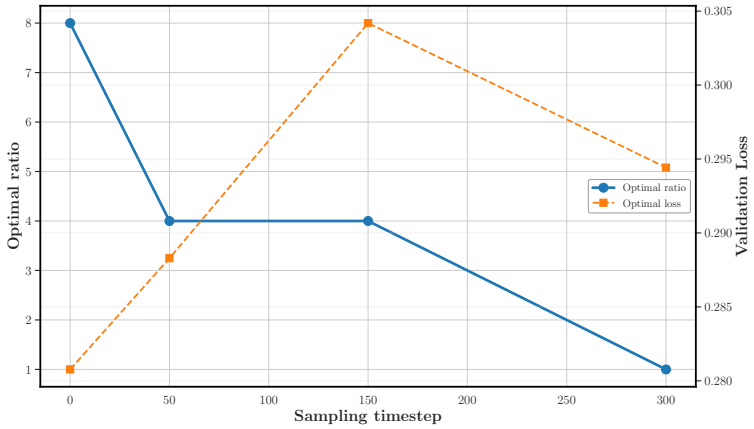
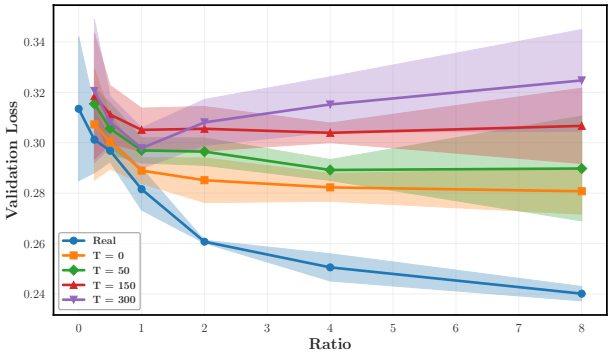
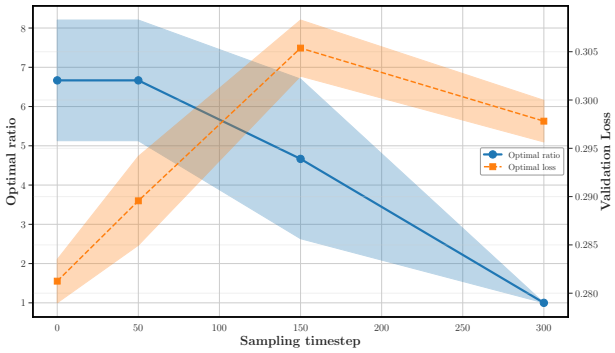


Figure 8: Optimal synthetic-to-real ratio (blue line) and the optimal validation loss (orange dashed line) as the distributional distance or equivalently the diffusion sampling timestep grows.

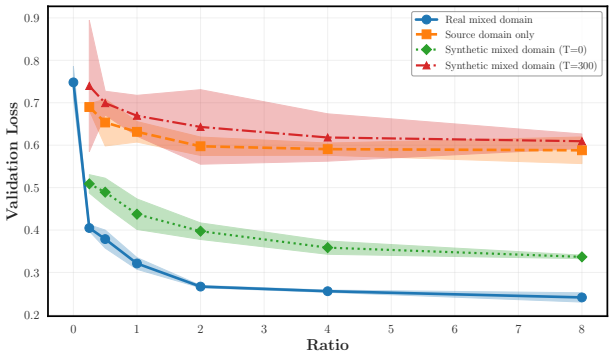
For reproducibility, we repeat each experiment using three different random seeds to account for variability introduced by stochastic elements in the training and sampling processes. The reported results in Figure 9 represent the mean performance across these runs, with corresponding confidence intervals to capture the variability. This approach ensures that our conclusions are not driven by a particular random initialization and provides a more robust estimate of model behavior.



(a) Validation loss vs. synthetic-to-real data ratio.



(b) Optimal choice of ratio and final validation loss.



(c) Validation loss vs ratio with domain shift.

Figure 9: (a) Validation loss vs. synthetic-to-real data ratio across different sampling timesteps, representing varying distributional distances. We observe a sharper U-turn effect as the distributional distance increases. (b) The optimal synthetic-to-real data ratio decreases as synthetic samples become noisier or deviate further from the true distribution. (c) Incorporating synthetic data improves out-of-domain generalization when the synthetic data distribution is closer to the target than the source. All experiments are repeated with three different seeds. The results align with those shown in the main text for single experimental runs.

I.4 Practical Insights

In this section, we investigate the effects of signal-to-noise ratio, i.e., heterogeneity, and the regularity of the problem. As shown in Figure 10, varying the noise level across different values of r exhibits a pattern consistent with our observations in Section 6. We again find that a 1:2 ratio of real to synthetic data performs well across these scenarios. While changes in the regularity of the objectives (i.e., r) influence the scale of the test error, the overall behavior remains consistent.

Similarly, under domain shift (see Figure 11), the effect of the signal-to-noise ratio aligns with the in-domain behavior reported in Section 6. Specifically, higher heterogeneity necessitates more careful selection of the real-to-synthetic ratio, as higher ratios can degrade performance when the distributional distance from the target domain is large.

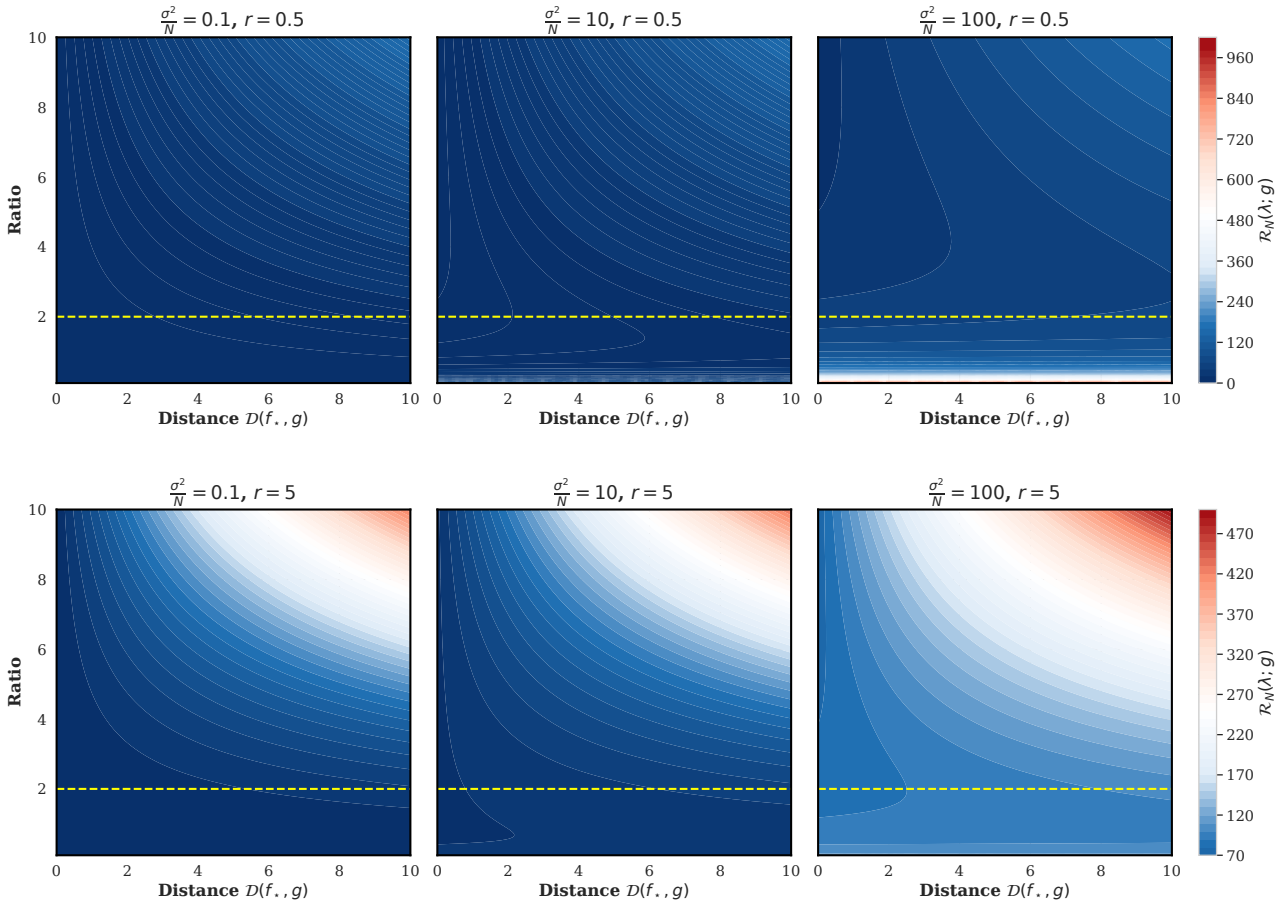


Figure 10: Effect of signal-to-noise ratio on the choice of optimal synthetic-to-real data ratio, across two different values of $r \in \{0.5, 5\}$.

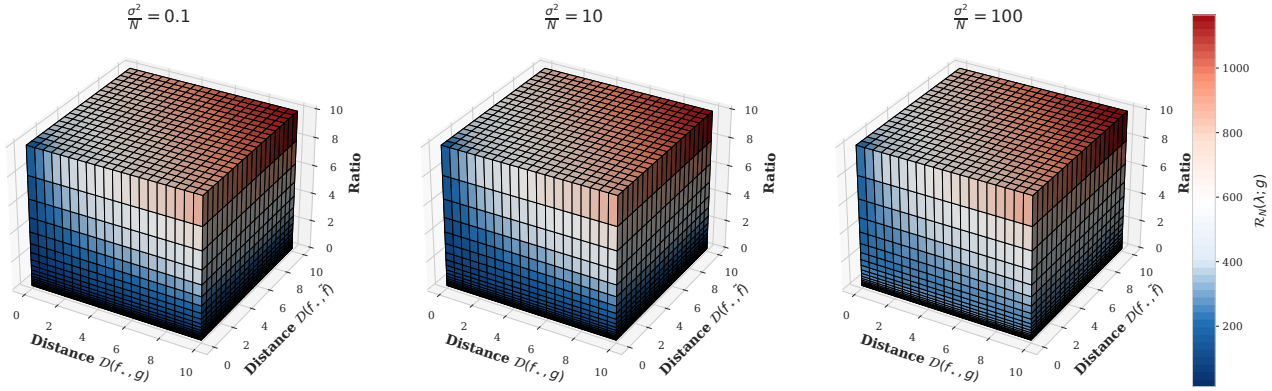


Figure 11: Effect of signal-to-noise ratio on the generalization error. All the plots are with $r = 1, \mu_{\max} = 1$.

J EXPANDED RELATED WORK

Synthetic Data The rapid advancement of generative models has significantly improved data generation quality, making it increasingly difficult to distinguish between synthetic and real data. Many previous works (Zhang et al., 2015; Cortés et al., 2024; Shrivastava et al., 2017; Lee et al., 2024; Seib et al., 2020; Zhezherau and Yanockin, 2024) have demonstrated the effectiveness of synthetic data in enhancing the performance of deep learning methods and stabilizing training, both in supervised and unsupervised settings, through augmentation and various applications. Alemohammad et al. (2024a); Shumailov et al. (2024); Briesch et al. (2023) analyzed the effects of training generative models on synthetic data across multiple iterations, creating a self-consuming loop, and found that without sufficient fresh real data, model quality or diversity deteriorates over time, leading to model collapse. To further investigate this, Dohmatob et al. (2024) studied model collapse theoretically in the regression case. Similarly, Bertrand et al. (2024); Gerstgrasser et al. (2024) looked at iteratively training generative models on mixed datasets, concluding that stability is maintained if the initial model is accurate and the real data proportion is sufficiently high. Ferbach et al. (2024) expanded on this by investigating the impact of user-curated synthetic data on iterative retraining, framing it as an implicit preference optimization process and exploring its theoretical effects on model stability and quality. In contrast, our paper does not examine iterative learning but instead focuses on a single-step approach, framing synthetic data as a regularizer. Possibly closest to our work is Jain et al. (2024), where the authors use a weighted empirical risk minimization approach to integrate surrogate data, reducing test error even when unrelated to the original data. However, their work differs from ours in two main aspects: (1) they focus on the scaling law of the test error, while we aim to determine the optimal synthetic-to-real data ratio or regularizer weight; (2) they do not account for the distance between synthetic and real data distributions, while our work specifically provides a bound based on this difference.

Optimal Mixing and Model Collapse While the high-level U-shaped behavior of test error with respect to the proportion of synthetic data is intuitively expected, our contribution lies in deriving it formally via a distinct analysis route: (i) a tractable kernel-regression formulation yielding the optimal ratio, and (ii) a general stability-based bound incorporating Wasserstein distances, which to our knowledge has not appeared previously. Several concurrent or recent works study related phenomena from complementary perspectives. Zhang et al. (2023) focuses on self-consuming generative models and out-of-distribution generalization of mixture data; the setting differs from our mixed-data generalization analysis, which considers a single-step ERM scenario with a fixed pool of synthetic samples rather than recursive retraining. We highlight conceptual overlaps in studying the limits of synthetic data, but the assumptions differ substantially (e.g., their recursive training loop vs. our single-round ERM formulation). Garg et al. (2025) and He et al. (2025) both derive optimal mixing effects but rely on different technical frameworks (e.g., asymptotic kernel limits, hierarchies of synthetic corruption models). Crucially, these works do not provide general stability-based bounds, which is where our contribution is novel and complementary. We view these concurrent results as reinforcing the importance of the optimal-ratio question while highlighting the diversity of proof techniques that can be brought to bear on it.

Noise Injection, Data Augmentation, and Distributional Mismatch A natural question is how our framework relates to the classical literature on training with noise and data augmentation. Prior works (Bishop, 1995; Lin et al., 2022; Shen et al., 2023) mainly analyze perturbations of inputs or labels under the assumption that the augmented data is sampled from the same underlying distribution. In contrast, our setting explicitly considers a distinct synthetic distribution, with a potentially non-zero Wasserstein distance from the real data. This structural domain shift materially changes the generalization behavior and is the key modelling distinction of our work. Concretely, while Bishop (1995) shows that noise training induces Tikhonov regularisation, it does not derive an optimal mixing ratio nor account for distribution mismatch. Similarly, Lin et al. (2022) and Shen et al. (2023) study benign overfitting or domain adaptation but do not provide an explicit dependence on the distributional distance between sources. Among these, Shen et al. (2023) is the closest to our work. However, although that paper provides generalization bounds and discusses the bias–variance tradeoff, it focuses primarily on aligning multi-source data to the target distribution in the unsupervised setting, where the goal is to leverage additional sources of feature space and label them according to their similarity or dissimilarity to the source data. In contrast, we explicitly introduce a tradeoff that depends on the distributional distance and derive an explicit interior optimum for the mixing ratio. More broadly, while bias–variance ideas appear throughout the literature, our main contribution is a general algorithmic-stability bound that yields an analytic form for the tradeoff, explicit dependence on the Wasserstein distance, and applicability to general learning algorithms—not

only kernel methods or linear models. This allows us, for example, to show that the U-curve sharpens with distributional mismatch, a phenomenon that does not arise in classical noise-injection analyses.

Mixing Data Sources Recent work has investigated scaling laws for mixing data sources in pretraining objectives (Ye et al., 2025; Shukor et al., 2025; Hashimoto, 2021), showing how performance depends on the proportions of different datasets. These studies typically require either running a large suite of experiments with varying source mixtures or access to massive datasets that can be systematically subsampled to empirically fit the law. In contrast, our work is motivated by applications such as healthcare and medical imaging, where data scarcity makes such large-scale experimentation infeasible. Methodologically, our approach provides an a priori estimate of the optimal synthetic-to-real data ratio, eliminating the need to empirically fit a scaling law. From a theoretical perspective, our analysis also departs substantially from existing work: whereas prior studies rely on empirical findings and use theoretical insights primarily for justification, we adopt a stability-theoretic framework that enables a precise characterisation of the trade-off and yields an explicit formula for the optimal ratio that practitioners can apply directly. Finding the optimal mixture of data sources is also closely related to identifying which sources are most *useful* for downstream performance. For example, Thudi et al. (2025) approaches the problem via a bi-level optimisation framework rather than empirical scaling laws, and analyses how the benefits of their method evolve with increasing model size. Similarly, Firdoussi et al. (2025) applies data pruning based on a score function to select high-quality samples, thereby improving the effectiveness of mixed data. Our work differs from these approaches in that we do not attempt to select or filter the synthetic data based on its quality. Instead, we assume the synthetic data is given and focus on how to combine it with the available real data in an optimal way. This perspective is particularly relevant in practical scenarios where the synthetic data generation process is fixed or externally provided, and the key challenge lies in determining how best to exploit it in conjunction with scarce real data.

Domain Adaptation and Transfer Learning Recent research has explored the intersection of domain adaptation and synthetic data, showing how synthetic data can bridge the gap between source and target domains, thereby enhancing model transferability and generalization across tasks (Mullick et al., 2023; Peng et al., 2018; Imbusch et al., 2022; Shakeri et al., 2020). A major challenge in transfer learning is the distribution gap, and several studies address this by using synthetic data to fine-tune models, improving generalizability (Mishra et al., 2022; Kim et al., 2020; Sariyildiz et al., 2023). Gerace et al. (2022) propose synthetic data as a framework for modeling correlations between datasets, showing improvements in generalization when transferring learned features from source to target tasks. On a more theoretical level, several works connect domain adaptation to distributionally robust learning, demonstrating that adding unlabeled or labeled data improves generalization; these setups can be easily extended to include synthetic data (Saberi et al., 2024a,b; Wu et al., 2022; Hou et al., 2023). Our perspective differs in that we do not view synthetic data as a replacement or proxy for domain adaptation, but rather as a complementary source whose utility depends critically on how it is balanced with real data. This focus on the optimal ratio naturally connects to knowledge distillation (Hinton et al., 2015; Stanton et al., 2021; Menon et al., 2020; Busbridge et al., 2025), where the objective similarly involves balancing two heterogeneous sources of supervision: the labelled data (analogous to real samples) and the teacher's soft signal (analogous to synthetic information). Just as we provide an explicit characterisation of the optimal real-to-synthetic ratio, one can view distillation through the same lens of optimally weighting complementary signals, highlighting a deeper connection between synthetic data integration and distillation-based learning.

Generalization Bounds The first generalization bounds were based on characterizations of the hypothesis space's complexity, such as the VC dimension or Rademacher complexity (Bousquet et al., 2003; Vapnik, 2000; Shalev-Shwartz and Ben-David, 2014). However, due to their algorithm-independent nature, these bounds must hold even for the worst algorithm within a given hypothesis space, making them often inadequate for modern over-parameterized neural networks, where the complexity measure typically scales exponentially with the architecture's depth (Anthony and Bartlett, 2002; Zhang et al., 2017; Belkin et al., 2018). To address this issue, recent approaches focus on providing algorithm-dependent generalization bounds. The underlying intuition is that a hypothesis less dependent on the input dataset is less prone to overfitting and, therefore, generalizes better. Among the results building on this idea are bounds based on uniform stability (Bousquet and Elisseeff, 2002; Attia and Koren, 2022), differential privacy (Dwork and Roth, 2014), PAC-Bayesian bounds (Guedj, 2019; McAllester, 1999), information-theoretic bounds (Russo and Zou, 2020; Gálvez et al., 2020; Haghifam et al., 2020), and chained bounds (Clerico et al., 2022; Asadi et al., 2018). Our work mainly uses the previously

established ideas of generalization bounds for mixed of real and synthetic data when the synthetic data acts as regularizer. More related to our work and on the importance of regularization, Li and Zhang (2021) analyzes the generalization properties of fine-tuning in transfer learning and proposes a PAC-Bayes generalization bound, combining regularization and self-labeling. Mou et al. (2018) provides generalization guarantees for dropout training by bounding the error using offset Rademacher complexities, capturing data-dependent regularization and the effect of perturbation variance.

K NOTES ON THE CHOICE OF DISTANCE METRIC

In Section 6 we discussed possible ways of estimating the Wasserstein distance for practical scenarios. However, one setup that happens in many practical cases particularly in healthcare domain is heterogeneous tabular data. Indeed, Wasserstein implicitly assumes a meaningful ground metric on the input space, which can be problematic when the data contains a mixture of categorical, ordinal, and continuous features, since arbitrary encodings or feature scalings can dominate the geometry. One possible direction is to first obtain a more homogeneous representation of tabular inputs using an unsupervised or semisupervised embedding (e.g., contrastive encoders (Chen et al., 2020; Caron et al., 2021; Shidani et al., 2024) or variational embeddings (Kingma and Welling, 2014)) and then apply the Wasserstein-based bound in this learned feature space. Another direction is to replace Wasserstein with discrepancies better suited to tabular data. Total variation (TV) is appealing for categorical or mixed tabular features because it requires no ground metric and tightly controls differences in expectations. However, TV quickly becomes too coarse and especially hard to estimate. This motivates using alternatives like Maximum Mean Discrepancy (MMD) with characteristic kernels (robust to mixed feature types and statistically favorable in high dimension) Gretton et al. (2012). These alternatives avoid imposing a single global geometry and often yield discrepancy measures more faithful to the structure of tabular data. Systematically studying how our stability-based generalization bound behaves under TV or these alternative divergences is an interesting future direction.