Pseudo-Labeling for Unsupervised Domain Adaptation with Kernel GLMs

Nathan Weill
Department of IEOR
Columbia University
New York, NY 10027
nmw2142@columbia.edu

Kaizheng Wang
Department of IEOR and Data Science Institute
Columbia University
New York, NY 10027
kw2934@columbia.edu

Abstract

We propose a principled framework for unsupervised domain adaptation under covariate shift in kernel Generalized Linear Models (GLMs), encompassing kernelized linear, logistic, and Poisson regression with ridge regularization. Our goal is to minimize prediction error in the target domain by leveraging labeled source data and unlabeled target data, despite differences in covariate distributions. We partition the labeled source data into two batches: one for training a family of candidate models, and the other for building an imputation model. This imputation model generates pseudo-labels for the target data, enabling robust model selection. We establish non-asymptotic excess-risk bounds that characterize adaptation performance through an "effective labeled sample size," explicitly accounting for the unknown covariate shift. Experiments on synthetic and real datasets demonstrate consistent performance gains over source-only and importance weighting baselines.

1 Introduction

The standard machine learning framework relies on the fundamental assumption that training and test datasets are drawn from the same underlying distribution (Vapnik, 1999). However, in realistic deployment scenarios, this assumption is frequently violated. A canonical instance of distribution shift is *covariate shift* (Shimodaira, 2000; Ben-David et al., 2010), where the marginal distribution of features varies between the source and target domains, while the conditional distribution of the output label remains stable. This phenomenon is pervasive in high-stakes applications, from personalized medicine to computer vision. It is well-documented that covariate shift can severely degrade predictive performance (Koh et al., 2021).

In many applications, obtaining labels for the target population is prohibitively expensive, whereas collecting unlabeled covariates is straightforward. This motivates the study of unsupervised domain adaptation (UDA), where the learning algorithm must generalize to the target domain using labeled source data and only unlabeled target covariates. We investigate UDA for ridge-regularized Generalized Linear Models (GLMs) and their kernel extensions. In this regime, simply applying a model trained on the source is often suboptimal, even if the model is well-specified. Standard learning algorithms tune regularization hyperparameters to balance bias and variance on the *source* distribution, ignoring regions of the covariate space where the target density is high but source data is scarce. Consequently, achieving optimal target performance requires an adaptive strategy that leverages unlabeled target data to recalibrate the estimator.

Contributions We address this challenge by developing a pseudo-labeling framework tailored for kernel GLMs. The approach enables model selection for the target domain using only unlabeled data. Our contributions are two-fold:

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

- **Methodology:** We propose a procedure that partitions the labeled source data into two batches. The first batch is used to train a family of candidate models, while the second trains an imputation model to generate pseudo-labels for the target covariates. These pseudo-labels serve as a proxy for ground truth, allowing us to select the candidate model that minimizes the estimated target risk. We provide precise guidance for tuning the imputation model to optimize the final selection.
- **Theory:** We derive non-asymptotic excess risk bounds for ridge-regularized kernel GLMs in the UDA setting. Our bounds quantify the information transfer via an "effective labeled sample size" that accounts for the discrepancy between source and target spectral properties. This theoretical result demonstrates that our method automatically adapts to the unknown covariate shift.

Related Work The theoretical landscape of transfer learning largely focuses on settings where the learner has access to labeled target data or a known target covariate distribution (Ben-David et al., 2010; Kpotufe and Martinet, 2021). Such assumptions simplify adaptation by allowing for direct risk estimation or precise density ratio calculation. In contrast, we address the strictly unsupervised regime, where the target distribution is unknown and observed only through a finite set of unlabeled covariates. This necessitates inferring the structure of the shift and adapting the estimator solely from the geometry of the unlabeled data, without the guidance of target labels.

The necessity of adaptation under covariate shift depends on the modeling regime. In the *parametric* setting, if the model is well-specified, standard Maximum Likelihood Estimation (MLE) on the source is asymptotically efficient for the target, rendering adaptation unnecessary (Ge et al., 2024). However, if the parametric model is mis-specified, the source MLE is biased for the target; the standard remedy is importance weighting (Shimodaira, 2000; Sugiyama et al., 2008) or its doubly robust variants utilizing pseudo-outcomes (Liu et al., 2023). While statistically consistent, these methods often suffer from high variance and the difficulty of estimating density ratios in high dimensions.

In the *nonparametric* regime we consider, the situation differs. Even if the model is well-specified (e.g., the true function lies in an RKHS), regularization is required to ensure finite-sample generalization. Recent works have established theoretical guarantees for reweighting-based adaptation in kernel ridge regression (KRR) (Ma et al., 2023) and kernel methods with more general losses (Feng et al., 2023). However, these approaches remain dependent on the existence and accurate estimation of density ratios. An alternative that avoids density estimation is to adapt the regularization directly. Wang (2026) addressed this for KRR using pseudo-labels to estimate the target error for selecting the optimal regularization. Our work operates in this same nonparametric regime and deliberately keeps the split–impute–select template of Wang (2026); our contribution is to prove that it remains valid for kernel GLMs (logistic, Poisson, ...), a case left open there, where the estimator has no closed form and the analysis of the KRR proof no longer applies (see Table 1 for a detailed comparison).

The strategy of filling missing labels with model predictions has driven empirical success in deep learning (Lee, 2013; Sohn et al., 2020). Theoretical analysis of pseudo-labeling has mostly focused on classification under cluster assumptions or margin conditions (Liu et al., 2021; Wei et al., 2021; Cai et al., 2021), which do not translate to the continuous responses covered by our GLM framework.

2 Problem setup

2.1 Generalized linear models under covariate shift

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ and $\{(\boldsymbol{x}_{0i}, y_{0i})\}_{i=1}^{n_0}$ denote samples drawn from the source and target distributions, respectively. Here, $\boldsymbol{x}_i, \boldsymbol{x}_{0i} \in \mathcal{X}$ are covariates in a compact metric space $\mathcal{X}$, and $y_i, y_{0i} \in \mathbb{R}$ are real-valued responses. In the Unsupervised Domain Adaptation (UDA) setting, the target labels $\{y_{0i}\}_{i=1}^{n_0}$ are unobserved. Our goal is to learn a predictive model using the labeled source data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ and the unlabeled target covariates $\{\boldsymbol{x}_{0i}\}_{i=1}^{n_0}$ that achieves low error on the target population.

We assume that both domains share a common Generalized Linear Model (GLM) (Nelder and Wedderburn, 1972) in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{F}$. Specifically, conditioned on $\boldsymbol{x}$, the response y is drawn from an exponential family distribution whose density function has the form:

$$p(y|\boldsymbol{x}) = h(y, \tau) \exp\left(\frac{y f^*(\boldsymbol{x}) - a(f^*(\boldsymbol{x}))}{d(\tau)}\right), \tag{2.1}$$

where $a : \mathbb{R} \to \mathbb{R}$ is a known, strictly convex log-partition function; h and d are known functions; $\tau > 0$ is a (possibly unknown) dispersion parameter; and f^* belongs to the class $\mathcal{F}$. The conditional mean of the response is given by the canonical link:

$$\mathbb{E}[y \mid \boldsymbol{x}] = a'(f^*(\boldsymbol{x})). \tag{2.2}$$

Up to an affine transform, the corresponding negative log-likelihood function is:

$$\ell(f(\boldsymbol{x}), y) := a(f(\boldsymbol{x})) - yf(\boldsymbol{x}), \quad \forall f \in \mathcal{F}. \tag{2.3}$$

Classical examples include *linear regression* [$a(u) = u^2/2$], *logistic regression* [$a(u) = \log(1+e^u)$], and *Poisson regression* [$a(u) = e^u$].

The RKHS $\mathcal{F}$ is generated by a continuous, symmetric, positive semi-definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. By the Moore-Aronszajn Theorem (Aronszajn, 1950), there exists a feature space (Hilbert space) $\mathbb{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ and a mapping $\Phi : \mathcal{X} \to \mathbb{H}$ such that $K(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle_{\mathbb{H}}$ for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. The RKHS is defined as the class of linear functionals on the feature space:

$$\mathcal{F} = \{ f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R} \mid f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle \Phi(\boldsymbol{x}), \boldsymbol{\theta} \rangle_{\mathbb{H}}, \, \boldsymbol{\theta} \in \mathbb{H} \} . \tag{2.4}$$

Equipped with the norm $\|f_{\boldsymbol{\theta}}\|_{\mathcal{F}} := \|\boldsymbol{\theta}\|_{\mathbb{H}}$, $\mathcal{F}$ is a Hilbert space isometric to $\mathbb{H}$.

We adopt the following standard assumptions for the UDA setting:

Assumption 2.1 (GLM in RKHS and Random Design)**.** *The source dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ and target dataset $\mathcal{D}_0 = \{(\boldsymbol{x}_{0i}, y_{0i})\}_{i=1}^{n_0}$ are independent, with samples drawn i.i.d. within each domain.*

- ***Covariate Shift:*** *The covariate distributions for source and target are $\mathcal{P}$ and $\mathcal{Q}$, respectively, with $\mathrm{supp}(\mathcal{Q}) \subseteq \mathrm{supp}(\mathcal{P})$. The second-moment operators $\boldsymbol{\Sigma} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}}[\Phi(\boldsymbol{x}) \otimes \Phi(\boldsymbol{x})]$ and $\boldsymbol{\Sigma}_0 = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{Q}}[\Phi(\boldsymbol{x}) \otimes \Phi(\boldsymbol{x})]$ are trace-class.*
- ***Well-Specified Model:*** *There exists a true function $f^* = f_{\boldsymbol{\theta}^*} \in \mathcal{F}$ that generates the responses in both domains according to the distribution in* (2.1)*.*
- ***Finite Noise Variance:*** *Define noise variables $\varepsilon_i = a'(f^*(\boldsymbol{x}_i)) - y_i$ and $\varepsilon_{0i} = a'(f^*(\boldsymbol{x}_{0i})) - y_{0i}$. Conditioned on the covariates, $\{\varepsilon_i\}_{i=1}^n$ and $\{\varepsilon_{0i}\}_{i=1}^{n_0}$ have finite second moments.*

Assumption 2.1 is a *pure* covariate shift assumption: the same conditional law $p(y \mid \boldsymbol{x})$ holds in both domains. Our guarantees do not cover conditional or concept shift, under which source labels need not carry information about the target conditional and the pseudo-label risk proxy may be biased; this restriction is shared by the covariate-shift literature we build on (Shimodaira, 2000; Ma et al., 2023; Wang, 2026).

2.2 Ridge regularization for kernel GLMs

For any predictive model $f \in \mathcal{F}$, we evaluate its performance via the expected negative log-likelihood on the target population:

$$R(f) = \mathbb{E}_{(\boldsymbol{x},y) \sim \mathcal{Q}}[a(f(\boldsymbol{x})) - yf(\boldsymbol{x})] = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{Q}}[a(f(\boldsymbol{x})) - f(\boldsymbol{x})a'(f^*(\boldsymbol{x}))]. \tag{2.5}$$

Under Assumption 2.1, $R(f)$ is minimized by f^*. We define the *excess risk* as:

$$\mathcal{R}(f) = R(f) - R(f^*) = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{Q}}\Big[a(f(\boldsymbol{x})) - a(f^*(\boldsymbol{x})) - a'(f^*(\boldsymbol{x}))(f(\boldsymbol{x}) - f^*(\boldsymbol{x}))\Big]. \tag{2.6}$$

This quantity, which is the expected Bregman divergence defined by a, measures the discrepancy between the predicted conditional distribution and the truth.

A standard approach to estimating f^* is to minimize the regularized empirical risk on the source data:

$$\hat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n (a(f(\boldsymbol{x}_i)) - y_i f(\boldsymbol{x}_i)) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2 \right\}, \tag{2.7}$$

where $\lambda > 0$ is a regularization hyperparameter balancing data fit and smoothness. By the Representer Theorem (Schölkopf et al., 2001), the solution lies in the span of the source data:

$\hat{f}_\lambda(\cdot) = \sum_{i=1}^n \hat{\alpha}_i K(\boldsymbol{x}_i, \cdot)$. The coefficient vector $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^n$ is the solution to the convex optimization problem:

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \left\{ \frac{1}{n} \left(\mathbf{1}^\top \boldsymbol{a}(\boldsymbol{\alpha}) - \boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{\alpha} \right) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} \right\}, \tag{2.8}$$

where $\boldsymbol{K}_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, and $\boldsymbol{a}(\boldsymbol{\alpha}) = (a([\boldsymbol{K}\boldsymbol{\alpha}]_1), \ldots, a([\boldsymbol{K}\boldsymbol{\alpha}]_n))^\top$.

For source-only learning, λ is typically selected to minimize the source validation error. However, under covariate shift, an optimal penalty for the source may be suboptimal for the target, particularly if the target distribution is concentrated in directions where the source provides weak coverage. Ideally, if we had access to target labels $\{y_{0i}\}_{i=1}^{n_0}$, we would select the parameter λ minimizing the empirical target risk:

$$\frac{1}{n_0} \sum_{i=1}^{n_0} \left(a(\hat{f}_\lambda(\boldsymbol{x}_{0i})) - y_{0i} \hat{f}_\lambda(\boldsymbol{x}_{0i}) \right). \tag{2.9}$$

Due to missing labels, we require a strategy to estimate the target performance using only the unlabeled covariates.

3 Methodology

We propose an adaptive method for unsupervised domain adaptation with kernel ridge GLM (KR-GLM). The core idea is to construct a computable proxy for the empirical risk (2.9). Note that the conditional mean response $\mathbb{E}(y_{0i}|\boldsymbol{x}_{0i}) = a'(f^*(\boldsymbol{x}_{0i}))$ is given by the true model f^*. We will build an auxiliary "imputation model" $\tilde{f}$ trained on part of the labeled source data and then use the estimated conditional mean $\tilde{y}_{0i} = a'(\tilde{f}(\boldsymbol{x}_{0i}))$ to replace y_{0i} in (2.9). The full procedure is detailed in Algorithm 1.

Algorithm 1 Pseudo-Labeling for Kernel GLM

1: **Input**: Source data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, unlabeled target dataset $\mathcal{D}_0 = \{\boldsymbol{x}_{0i}\}_{i=1}^{n_0}$, training split size $n_1 \in [n]$, candidate penalty grid $\Lambda \subset (0, \infty)$, imputation penalty $\tilde{\lambda} > 0$.

2: **Step 1 (Data Splitting)**. Randomly partition the source data into two disjoint subsets: $\mathcal{D}_1$ of size n_1 (for candidate training) and $\mathcal{D}_2$ of size $n_2 = n - n_1$ (for imputation).

3: **Step 2 (Training)**.

4: (a) **Candidate Models**: For each $\lambda \in \Lambda$, train a candidate $\hat{f}_\lambda$ on $\mathcal{D}_1$ by solving the program

$$\min_{f \in \mathcal{F}} \left\{ \frac{1}{n_1} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_1} (a(f(\boldsymbol{x})) - y f(\boldsymbol{x})) + \frac{\lambda}{2} \|f\|_{\mathcal{F}}^2 \right\}.$$

5: (b) **Imputation Model**: Train the imputation model $\tilde{f}$ on $\mathcal{D}_2$ with penalty $\tilde{\lambda}$ by solving the program

$$\min_{f \in \mathcal{F}} \left\{ \frac{1}{n_2} \sum_{(\boldsymbol{x}, y) \in \mathcal{D}_2} (a(f(\boldsymbol{x})) - y f(\boldsymbol{x})) + \frac{\tilde{\lambda}}{2} \|f\|_{\mathcal{F}}^2 \right\}.$$

6: **Step 3 (Pseudo-Labeling)**. Generate pseudo-labels for the target data using the conditional mean of the imputation model:

$$\tilde{y}_{0i} = a'(\tilde{f}(\boldsymbol{x}_{0i})), \quad \text{for } i = 1, \ldots, n_0.$$

7: **Step 4 (Model Selection)**. Select the candidate $\hat{\lambda} \in \Lambda$ that minimizes the pseudo-target risk:

$$\frac{1}{n_0} \sum_{i=1}^{n_0} \left(a(\hat{f}_\lambda(\boldsymbol{x}_{0i})) - \tilde{y}_{0i} \hat{f}_\lambda(\boldsymbol{x}_{0i}) \right). \tag{3.1}$$

8: **Output**: The selected model $\hat{f} = \hat{f}_{\hat{\lambda}}$.

Soft vs. hard labeling. For classification tasks such as logistic regression, our method utilizes "soft" pseudo-labels $\tilde{y}_{0i} = 1/(1 + e^{-\widetilde{f}(\boldsymbol{x}_{0i})}) \in (0, 1)$—representing the estimated conditional probability—rather than "hard" labels $\mathbf{1}(\widetilde{f}(\boldsymbol{x}_{0i}) \geq 0) \in \{0, 1\}$. This distinction is critical. While hard pseudo-labeling is common in semi-supervised learning to enforce low-entropy decision boundaries (Lee, 2013; Sohn et al., 2020; Liu et al., 2021; Wei et al., 2021; Cai et al., 2021), our objective is to minimize the negative log-likelihood to recover the true conditional distribution. Using hard labels would destroy the calibration information necessary for minimizing the log-likelihood, introducing severe bias into the risk proxy (3.1) by forcing candidate models to fit extreme probabilities (0 or 1) rather than the true underlying uncertainty.

Implementation and tuning strategy. The procedure separates model training (Step 2a) from validation (Steps 2b–4): the candidates are scored on the target covariates $\mathcal{D}_0$ with pseudo-labels produced by an imputer trained on $\mathcal{D}_2$, which is independent of the data $\mathcal{D}_1$ used to train them. The computational cost is dominated by solving the kernel GLM optimization, which is efficient for standard losses. Our theoretical analysis (presented in Section 4) provides specific guidance on parameter choices:

- **Split ratio**: We typically set $n_1 = \lfloor n/2 \rfloor$, using half the source data for training candidates and half for imputation. This balanced split is standard in split-sample validation methods.
- **Candidate grid (Λ)**: The grid should be sufficiently dense to cover the bias-variance spectrum. We typically set Λ as a geometric sequence ranging from $O(n^{-1})$ (low bias, high variance) to $O(1)$ (high bias, low variance), with a cardinality of roughly $O(\log n)$. In practice, we construct it heuristically from source data only: we inspect the source regularization curve, choose a range spanning its high-variance and high-bias regimes, and extend it by about one order of magnitude beyond the low-penalty end.
- **Imputation penalty ($\tilde{\lambda}$)**: This is the critical hyperparameter. Unlike standard supervised learning, where $\tilde{\lambda}$ would be tuned to minimize source prediction error, here it must be chosen to optimize the *model selection* capability of the pseudo-labels. Our theory in Appendix B.1 suggests setting $\tilde{\lambda} \asymp n^{-1}$ up to logarithmic factors. This rate typically corresponds to "undersmoothing" the imputation model—prioritizing low bias in the pseudo-labels over low variance—which is crucial for accurately ranking the candidate models on the target domain. Concretely, we use the rule $\tilde{\lambda} = \min \Lambda$, which requires no target labels and no knowledge of the shift or of problem-dependent constants. The ablation in Appendix H.3 confirms the need for undersmoothing and shows that the final risk is insensitive to the exact value of $\tilde{\lambda}$ within the undersmoothed range.

4 Theoretical guarantees on the adaptivity

This section develops theoretical guarantees for Algorithm 1. We show that the resulting estimator adapts both to the target distribution's structure and to the unknown covariate shift. Our analysis introduces an effective sample size that quantifies how informative the labeled source data are for the target regression task.

4.1 Assumptions

Our analysis of kernel ridge GLM under covariate shift (Assumption 2.1) necessitates only mild regularity conditions, that one can verify are satisfied in most common scenarios, as discussed in Appendix A.3.

First, we need some assumptions on the kernel function K. Mercer's theorem (see for instance Wainwright (2019)) guarantees the existence of an orthonormal eigen-expansion of the PSD kernel with respect to a measured function class when it is continuous and satisfies the Hilbert-Schmidt condition.

Assumption 4.1 (Kernel regularity)**.** *We assume Mercer's theorem holds under both source and target distributions, impose a uniform boundedness assumption on the eigenfunctions, and a polynomial decay of the eigenvalues:*

- *For any $\boldsymbol{x}, \boldsymbol{z} \in \mathcal{X}$, $K(\boldsymbol{x}, \boldsymbol{z})$ can be expanded as*

$$\sum_{j=1}^{\infty} \mu_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{z}) = \sum_{j=1}^{\infty} \mu_{0j} \phi_{0j}(\boldsymbol{x}) \phi_{0j}(\boldsymbol{z})$$

  where $(\mu_j)_{j=1}^{\infty}$ and $(\mu_{0j})_{j=1}^{\infty}$ are sequences of positive eigenvalues [1]*, $(\phi_j)_{j=1}^{\infty}$ is an orthonormal sequence of eigenfunctions in $L^2(\mathcal{X}; \mathcal{P})$, and $(\phi_{0j})_{j=1}^{\infty}$ is an orthonormal sequence of eigenfunctions in $L^2(\mathcal{X}; \mathcal{Q})$. The convergence of the infinite series holds absolutely and uniformly.*

- $\exists M_1 < \infty, \forall j, \|\phi_j\|_{\infty} \leq M_1, \|\phi_{0j}\|_{\infty} \leq M_1$.
- $\exists A, A' > 0$, $\exists \alpha \geq 1$, $\forall j \geq 1, \mu_j \leq A j^{-2}$, $\mu_{0j} \leq A' j^{-2\alpha}$.

These assumptions hold in particular for the linear and polynomial kernels on compact domains, as well as for the first-order Sobolev kernel.

We also assume bounded noise and signal. The noise is only assumed sub-exponential instead of sub-Gaussian to include Poisson regression in our analysis.

Assumption 4.2 (Sub-exponential noise)**.** *Conditioned on $\boldsymbol{X}$, the noise vector $\boldsymbol{\varepsilon}$ satisfies $\|\boldsymbol{\varepsilon}\|_{\psi_1} \leq \sigma < \infty$.*

Assumption 4.3 (Bounded signal)**.** $\|f^*\|_{\mathcal{F}} \leq R < \infty$.

Remark 4.1. Assumptions 4.1 and 4.3 imply that there exists a constant $C < \infty$, such that for any $\boldsymbol{x} \in \operatorname{supp}(\mathcal{P})$, $|f^*(\boldsymbol{x})| \leq C$.

Finally, we assume some smoothness and convexity conditions for the log-partition function:

Assumption 4.4 (Local strong convexity and boundedness)**.** *We assume that the log-partition function a is strictly convex, smooth, and strongly convex in compact regions: for any $\rho > 0$, we have $k_\rho := \inf_{|z| \leq (\rho+1)C} a''(z) > 0$. Define $k = k_1$. Moreover, a'' is assumed to be upper bounded and Lipschitz on compact regions: let $K_\rho := \sup_{|z| \leq (\rho+1)C} a''(z) < \infty$, $L_\rho := \sup_{|z| \leq (\rho+1)C} a'''(z) < \infty$, and define $K = K_1$, $L = L_1$. Then, $\forall |z| \leq 2C, k \leq a''(z) \leq K$ and $\forall |z_1|, |z_2| \leq 2C, |a''(z_1) - a''(z_2)| \leq L\,|z_1 - z_2|$.*

4.2 Main results

Our main theorem shows that our method adapts to the unknown covariate shift.

Theorem 4.1 (Excess risk)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold. Define for $\mu^2 = M_1^6 \sigma^2$*

$$n_{\mathit{eff}} = \sup \left\{ t \leq n \;\middle|\; t\boldsymbol{\Sigma}_0 \preceq n\boldsymbol{\Sigma} + \frac{\mu^2}{k} \boldsymbol{I} \right\}. \tag{4.1}$$

Choose any $\delta \in (0, 1/e]$. Consider Algorithm 1 with $n_1 = n/2$, $\tilde{\lambda} = \mu^2 \log^7(n) \log(n_0/\delta)/n$, and

$$\Lambda = \left\{ 2^{j-1} \mu^2 / n : 1 \leq j \leq \lceil \log_2 n \rceil + 1 \right\}.$$

Then there exists a function ζ polylogarithmic in (n, n_0, δ^{-1}) multiplied by a polynomial in $\frac{K}{k}$, such that with probability at least $1 - \delta$:

$$R(\hat{f}) - R(f^*) \leq \zeta \cdot \inf_{\rho > 0} \left\{ \rho R^2 + \frac{\sigma^2}{n_{\mathit{eff}}} \sum_{j=1}^{\infty} \frac{\mu_{0j}}{k\mu_{0j} + \rho} \right\} + \zeta \cdot R^2 (1 + L^2) \mu^2 \left(\frac{1}{n_{\mathit{eff}}} + \frac{1}{n_0} \right). \tag{4.2}$$

The proof is deferred to Appendix C.3.

Remark 4.2 (Interpretation of Effective Sample Size)**.** The quantity n_{eff} defined in Eq. (4.1) serves as a data-dependent measure of the covariate shift. It quantifies the extent to which the source covariance $\boldsymbol{\Sigma}$ "covers" the target covariance $\boldsymbol{\Sigma}_0$. Crucially, this alignment is not merely a check of input support, but is measured through the *geometry of the kernel*: the covariance inequality dictates how well the source data captures variance along the principal directions of the RKHS relative to the target distribution.

[1] We order them as $\mu_1 \geq \mu_2 \geq \cdots > 0$ and $\mu_{01} \geq \mu_{02} \geq \cdots > 0$. Note that they are also the eigenvalues of the second moment operators $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ respectively.

- If the source dominates the target (i.e., $\mathbf{\Sigma}_0 \preceq B\mathbf{\Sigma}$), we have $n_{\text{eff}} \geq n/B$, recovering standard fast rates.
- If the target $\mathbf{\Sigma}_0$ has high energy in eigen-directions where the source $\mathbf{\Sigma}$ decays rapidly (a mismatch in kernel geometry), $n_{\text{eff}} \ll n$, reflecting the difficulty of adaptation. For instance, consider the first-order Sobolev space with $\mathcal{P} = \mathcal{U}[0,1]$ and $\mathcal{Q} = \delta_{0.5}$. It is proved in Section B.2 of Wang (2026) that $n_{\text{eff}} \asymp \sqrt{n}$.

The infimum over ρ in (4.2) captures the optimal bias-variance tradeoff for KRGLM on n_{eff} labeled target samples: the term ρR^2 is the regularization bias, which increases with ρ, while $\frac{\sigma^2}{n_{\text{eff}}} \sum_{j=1}^{\infty} \frac{\mu_{0j}}{k\mu_{0j}+\rho}$ is the variance, which decreases with ρ and reflects how well the target eigendirections are captured given n_{eff} effective samples. The infimum over ρ selects the regularization level that optimally balances these two competing terms, yielding the same rate as if one had direct access to n_{eff} labeled target samples. The second term in (4.2) reflects the extra error incurred by the imperfect pseudo-labels. The following corollary shows that this overhead is negligible compared to the first term in our non-parametric setting.

Corollary 4.1 (Convergence Rate with Effective Sample Size)**.** *Given the polynomial spectral decay assumption $\mu_{0j} \leq Aj^{-2\alpha}$, and provided n and $n_0/\log n$ are sufficiently large, the excess risk satisfies with probability at least $1 - n^{-1}$:*

$$R(\hat{f}) - R(f^*) \lesssim n_{\mathit{eff}}^{-\frac{2\alpha}{2\alpha+1}} + n_0^{-1}. \tag{4.3}$$

The $\lesssim$ only hides polylogarithmic factors in n and n_0, and constants depending on the GLM parameters. The proof is deferred to Appendix C.4.

Remark 4.3 (Required number of target covariates)**.** The pseudo-label term n_0^{-1} is negligible as soon as $n_0 \gtrsim n_{\text{eff}}^{2\alpha/(2\alpha+1)}$. This requirement is sublinear in n_{eff}, hence in n, and it moves in the opposite direction to the difficulty of the problem: a harder shift (smaller n_{eff}) slows the estimation rate and *weakens* the demand on n_0. For $\alpha = 1$ and $n = 10^4$, a hard shift with $n_{\text{eff}} \asymp \sqrt{n}$ only needs $n_0 \gtrsim 22$, whereas a mild shift with $n_{\text{eff}} \asymp n$ needs $n_0 \gtrsim 464$. The constraint thus binds for mild shifts with scarce target covariates (e.g., a rare subpopulation), where the bound degrades to $O(n_0^{-1})$ whatever the amount of source data; see Appendix C.5.

Remark 4.4 (Optimality and adaptivity)**.** The bound established in Corollary 4.1 is sharp. It is easily seen in the bounded density ratio regime $\mathrm{d}\mathcal{Q}/\mathrm{d}\mathcal{P} \leq B$ for some constant $B \geq 1$, implying $n_{\text{eff}} \geq n/B$. In this setting, we recover the rate $(n/B)^{-\frac{2\alpha}{2\alpha+1}}$, which matches the minimax lower bound for KRR with square loss derived in Theorem 2 of Ma et al. (2023). While their lower bound holds for estimators with full knowledge of $\mathcal{Q}$ and B, our Algorithm 1 achieves this rate without any prior information on the target distribution $\mathcal{Q}$ or the shift magnitude B, demonstrating its adaptivity.

Proof idea A comprehensive proof sketch is presented in Appendix B; here we summarize the main idea. First, we establish an *in-sample oracle inequality for model selection with pseudo-labels under the GLM loss* (Theorem B.1). This result quantifies how the imputation error affects the selected model and is fully generic, independent of the RKHS structure. It then remains to control the quantities appearing in this oracle inequality in the kernel GLM setting. Since KRGLM has no closed-form solution, we Taylor-expand the gradient of the regularized empirical loss around f^*, obtaining a bias–variance–perturbation decomposition. The perturbation term is the main non-square-loss difficulty: the Hessian depends on the estimator itself, so it can only be controlled after a *localization* step places the estimator in a strong-convexity neighborhood of f^*. The perturbation then reduces to pointwise deviations, yielding the *pointwise convergence rate* of Theorem B.2 as a key intermediate bound for the oracle-inequality terms. Combining these results yields Theorem 4.1.

Comparison with Wang (2026). Table 1 compares our analysis with the KRR analysis of Wang (2026). The split–impute–select pipeline is intentionally inherited, and our n_{eff} reduces to theirs at $k = 1$; our contribution is to prove that this procedure remains valid beyond least squares, a case that Wang (2026) leaves open, as the analytical expressions of the KRR estimator are central to their proof. Without them, the perturbation term above can only be controlled once the estimator is localized where $a'' \geq k$, which is itself what must be established; our localization argument and pointwise rate break this circularity, and the Bregman structure of the GLM loss extends the pseudo-label oracle

Table 1: Comparison with Wang (2026). Both use the same split–impute–select pipeline.

	Wang (2026) – KRR	This paper – KRGLM
Effective size	$\sup\{t \le n : t\boldsymbol{\Sigma}_0 \preceq n\boldsymbol{\Sigma} + \mu^2 \boldsymbol{I}\}$	Same, with curvature: $t\boldsymbol{\Sigma}_0 \preceq n\boldsymbol{\Sigma} + \mu^2 \boldsymbol{I}/k$
Loss / task	Square loss; regression	GLM loss: linear, logistic, Poisson (regression, classification, counts)
Curvature	Quadratic everywhere ($a'' \equiv 1$)	Locally: $k \le a'' \le K$ on $[-2C, 2C]$, a'' L-Lipschitz
Noise	Sub-Gaussian	Sub-exponential (needed for Poisson)
Kernel	Trace-class $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_0$, bounded feature map; decay only in the corollary	Mercer expansion with bounded eigenfunctions in the main theorem, for pointwise control (strictly stronger)
Estimator	Closed form; exact bias + linear noise split	Implicit; extra perturbation term, as the Hessian depends on the estimator
Imputer error	Immediate from the closed form	Localization + pointwise rate (Theorem B.2)
Oracle ineq.	General classes, squared loss	Same generality, GLM excess risk via the Bregman structure (Theorem B.1)

inequality. While reusing the loss-agnostic parts of their analysis (e.g., Lemma C.5), we obtain the same rate for logistic and Poisson regression, substantially enlarging the scope of the method: many UDA problems, including both of our real-data tasks, are classification problems.

5 Numerical experiments

We conduct experiments on synthetic and real data to show the benefit of our approach over source-only and reweighting baselines. The code for implementing our experiments is available at `https://github.com/nathanweill/pseudolabel-uda-experiments`. In particular, we implemented an ad hoc KRGLM solver that can be of independent interest; it is explained in detail in Appendix H.1.

5.1 Synthetic data

For simulation experiments, our feature space $\mathcal{X}$ is the unit interval $[0, 1]$, and the kernel K is the first-order Sobolev kernel augmented with a constant, $K(u, v) = 1 + \min\{u, v\}$, whose RKHS contains f^* below (Example A.3). Let $f^*(x) = 1.5\cos(2\pi x)$, $\nu_0 = \mathcal{U}[0, 1/2]$ and $\nu_1 = \mathcal{U}[1/2, 1]$. For any even integer $n \ge 2$, we follow the procedure below:

- Let $B = n^{0.4}$, $\mathcal{P} = \frac{B}{B+1}\nu_0 + \frac{1}{B+1}\nu_1$ and $\mathcal{Q} = \frac{1}{B+1}\nu_0 + \frac{B}{B+1}\nu_1$. Generate n i.i.d. source samples $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \sim \mathcal{P}$ and $y_i \mid x_i \sim \text{Bernoulli}\left(\sigma(f^*(x_i))\right)$, where σ is the sigmoid function. Generate $n_0 = n$ i.i.d. unlabeled target samples $\mathcal{D}_0 = \{x_{0i}\}_{i=1}^{n_0}$ from $\mathcal{Q}$.
- Run Algorithm 1 with $n_1 = n/2, \Lambda = \left\{2^k/(10n) : k = 0, 1, \cdots, \lceil \log_2(10n) \rceil\right\}$, and $\widetilde{\lambda} = 1/(10n)$ to get an estimate $\widehat{f}$. Use the log-partition function associated to logistic regression $a(u) = \log(1 + e^u)$, which verifies $a'(u) = \sigma(u)$.

The 1.5 amplitude of f^* ensures that the source contains enough points near the decision boundary to learn decent pseudo-labels. The quantity B controls the strength of the covariate shift; letting it grow with n allows us to test larger shifts at larger sample sizes.

Our method selects a candidate model that minimizes the (target) risk estimate

$$f \mapsto \frac{1}{|\mathcal{D}_0|} \sum_{x \in \mathcal{D}_0} \left(a\left(f(x)\right) - f(x) a'(\tilde{f}(x)) \right)$$

based on pseudo-labels. We compare it with an oracle selection approach whose (target) risk estimate

$$f \mapsto \frac{1}{|\mathcal{D}_0|} \sum_{x \in \mathcal{D}_0} \left(a\left(f(x)\right) - f(x) a'(f^*(x)) \right)$$

is constructed from the true conditional probabilities, and a naive approach that uses the empirical (source) risk

$$f \mapsto \frac{1}{|\mathcal{D}_2|} \sum_{(x,y) \in \mathcal{D}_2} \left(a\left(f(x)\right) - y f(x) \right)$$

on $\mathcal{D}_2$. The latter is the standard hold-out validation method that ignores the covariate shift.

These three criteria are used only for selecting λ from the candidate grid. Final performance is evaluated separately on an independent target test set, consisting of 10,000 new samples from $\mathcal{Q}$. We then report the corresponding Monte Carlo estimate of the target excess risk, computed via Eq. (2.6).

For every $n \in \{4000, 8000, 16000, 32000\}$, we conduct 100 independent runs of the experiment. The results are shown in Figure 1a in Appendix H.2. The horizontal axis represents the source sample size n, and the vertical axis represents the average estimated target excess risk across the 100 trials.

First, the oracle is lowest, as expected; pseudo-labeling closely tracks it; and the naive method is uniformly higher, reflecting the cost of selecting λ under $\mathcal{P}$ while evaluating under $\mathcal{Q}$. Second, the approximately linear alignment of the data points on the log-log scale suggests a power-law decay of the target excess risk, of the form $O(n^{-\kappa})$ for some exponent $\kappa > 0$. Linear fits to the log-transformed data give $\kappa = 0.546(0.047)$, $0.523(0.051)$, and $0.439(0.043)$ for the pseudo-labeling, oracle, and naive methods, respectively, represented by the solid lines in the figure. The standard errors reported in parentheses were computed via a cluster bootstrap procedure (Cameron et al., 2008), resampling the 100 trials associated with each n over 10,000 iterations. Our proposed method achieves an error exponent that does not differ significantly from the oracle benchmark. Furthermore, our approach significantly outperforms the naive method, which suffers from poor source coverage of the target domain.

Varying the hardness of the shift. To test the adaptivity of Algorithm 1 directly in the regime $n_{\text{eff}} \ll n$, we repeat the experiment with $B = n^{\gamma}$ for $\gamma \in \{0.40, 0.45, \ldots, 0.65\}$, which corresponds to $n_{\text{eff}} \asymp n^{1-\gamma}$, with 100 trials for each n. Table 4 in Appendix H.2 reports the fitted exponents. As the shift becomes harder, naive source validation deteriorates until its excess risk is essentially flat at $\gamma = 0.65$, whereas pseudo-labeling keeps improving, without knowledge of B or n_{eff}: at $n = 32{,}000$ and $\gamma = 0.65$, the target excess risks are 0.0429 (pseudo-labeling), 0.1361 (naive) and 0.0265 (oracle). The fitted exponents of pseudo-labeling and oracle exceed the worst-case minimax exponent $2(1-\gamma)/3$ of Corollary 4.1 (with $\alpha = 1$); this should not be over-interpreted, as the slopes are fitted from four sample sizes only, and a fixed f^* can converge faster than the worst case without contradicting the lower bound. Mean risks, paired slope comparisons and plots for every γ are available in the code repository.

5.2 The Raisin dataset

The Raisin dataset[2] (Çınar et al., 2020) contains 900 instances of two raisin varieties with 7 numerical features. The task is binary classification. To introduce covariate shift, we follow a usual subsampling technique on the first covariate (Quiñonero-Candela et al., 2008): after scaling the data, for each sample $\boldsymbol{x}_i$, look at its first component $\boldsymbol{x}_i[0]$ and compute the rejection probability $p_i = \min(1, (\boldsymbol{x}_i[0] - c)^2/l)$, where $c = \min_i(\boldsymbol{x}_i[0])$ and $l > 0$ is a parameter to adjust the shift strength (the smaller it is, the larger the shift). Then sample i is assigned to the target out-of-distribution (OOD) set if a draw of the random variable Bernoulli(p_i) returns 1, otherwise, it is assigned to the in-distribution (ID) source set. In the rest of this experiment, we selected a shift level $l = 3$. Since the shift is induced by covariate-dependent subsampling, $p(y \mid \boldsymbol{x})$ is unchanged and this benchmark satisfies the pure covariate shift assumption of our theory.

In Appendix H.4, we show that among ridge logistic regression models trained on source only, the target-optimal regularization parameter is around 100 times larger than the source-optimal one. Our method allows us to do this target-aware model selection without seeing any target label. This phenomenon is illustrated in Figure 3.

We run our pseudo-labeling experiment in the same fashion as for the synthetic data (Section 5.1), by comparing it to a naive method that selects the best model on source validation data, and to an oracle method that uses the target labels for selection. We added comparisons against importance-weighting (IW) methods using two density-ratio estimation procedures: Kernel Mean Matching (Huang et al., 2006) and Kullback-Leibler Importance Estimation Procedure (Sugiyama et al., 2008). Since Feng et al. (2023), our closest competitor in the kernel covariate-shift setting, do not provide a target-label-free selection rule, we extended their KLIEP implementation with importance-weighted cross-validation (Sugiyama et al., 2007) for a fair UDA comparison on the Raisin benchmark.

[2] available at `https://archive.ics.uci.edu/dataset/850/raisin`

Table 2: Risk mean estimate across seeds with confidence interval (CI).

Method	Mean	95% CI
Naive	0.502	[0.463, 0.541]
KLIEP + IW-CV	0.478	[0.438, 0.518]
KMM + IW-CV	0.469	[0.442, 0.496]
Pseudo-labeling	0.428	[0.412, 0.444]
Oracle	0.373	[0.361, 0.385]

We run it for 100 different random seeds for significance and report the results in Table 2. As our objective is to estimate the conditional probability function, we evaluate performance with the empirical risk (i.e., the log-loss), a strictly proper scoring rule. For each seed, we first split the OOD data into training (for pseudo, density-ratio and oracle methods) and test set for evaluation purposes.

Given the limited sample size, we adopted a more robust model selection strategy: instead of a simple hold-out split, we employed repeated 2-fold cross-validation (see for instance Arlot and Celisse (2010); Kim (2009)) (6 repetitions) on the In-Distribution (ID) data. Candidate models were selected based on risk estimates averaged across all folds, and subsequently refitted on the full ID dataset using the chosen hyperparameters. This refit is a standard empirical variance-reduction step, used only here: our guarantees cover the output of Algorithm 1, trained on half of the source data, not the refitted model; it is applied identically to all methods, so it cannot create the reported gaps (the synthetic and Camelyon17 experiments run Algorithm 1 as stated, without refitting). Our pseudo-labeling approach consistently outperforms the other baselines; the higher variance of the importance-weighted methods reflects the practical instability of density-ratio estimation in finite samples. See Appendix H.4 for method details, hyperparameters, and further results.

5.3 The Camelyon17 medical images

We finally test the selection principle well outside the scope of our theory, on early stopping when fine-tuning a deep network on the Camelyon17 (Bandi et al., 2018) benchmark from WILDS (Koh et al., 2021). This experiment is *not* evidence for Theorem 4.1: the candidates are neither kernel GLMs nor ridge-penalized, and scanner and staining differences across hospitals plausibly induce conditional, not merely covariate, shift. It tests whether the selection principle—impute the conditional mean on target covariates, then rank candidates by pseudo-target risk—transfers to a modern workflow, here tumor detection in histopathological patches. A pre-trained DINOv2 ViT-B/14 (Oquab et al., 2023) backbone is fine-tuned end-to-end on labeled source data ($\approx$ 300k samples from hospitals 1–3), with the epoch number serving as the regularization parameter. Following Algorithm 1, an imputer is trained on a held-out source validation dataset ($\approx$ 34k samples) and used to generate soft pseudo-labels on an unlabeled target selection set ($\approx$ 42k samples from hospital 5); the selected checkpoints are evaluated on a held-out target test set ($\approx$ 42k samples). Across three training seeds, naive source validation selects late epochs (8, 18, 21), after the model has overfit to source-specific features, with a mean target log-loss of 2.09. Pseudo-labeling does not always recover the exact oracle epoch, but it consistently selects an early-stopping region and beats naive validation in every run, with a mean target log-loss of 0.82 (oracle: 0.43), i.e., a mean regret of 0.39 versus 1.66 for naive validation. Results are stable across target selection/test splits. Per-seed results and further details are in Appendix H.5.

6 Conclusion

We presented a principled framework for unsupervised domain adaptation in kernel GLMs, utilizing a split-and-fit strategy to perform target-aware model selection via pseudo-labeling. Our non-asymptotic excess risk bounds characterize adaptation through an "effective labeled sample size", and our experiments confirm consistent gains over source-only and reweighting baselines. Our pure covariate shift assumption remains a limitation, as real-world data typically exhibit a mixture of distribution shifts. Several future directions are worth exploring. Extending our framework beyond ridge regularization and RKHS estimators would broaden its applicability. Also, our analysis focuses on one-step imputation; analyzing iterative refinement, where pseudo-labels are updated over multiple rounds, could connect our bounds to the empirical success of self-training.

Acknowledgments and Disclosure of Funding

Funding. The research is supported by NSF grants DMS-2210907 and DMS-2515679, and a Data Science Institute seed grant SF-181 at Columbia University.

Competing interests. The authors declare no competing interests.

References

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

N Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68 (3):337–404, 1950.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, May 2010. ISSN 0885-6125, 1573-0565. doi: 10.1007/s10994-009-5152-4.

Tianle Cai, Ruiqi Gao, Jason Lee, and Qi Lei. A theory of label propagation for subpopulation shift. In *International Conference on Machine Learning*, pages 1170–1182. PMLR, 2021.

A. C. Cameron, J. B. Gelbach, and D. L. Miller. Bootstrap-based improvements for inference with clustered errors. *The review of economics and statistics*, 90:414–427, 2008.

Benjamin Charlier, Jean Feydy, Joan Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

İlkay Çınar, Murat Köklü, and Şakir Taşdemir. Raisin. UCI Machine Learning Repository, 2020.

Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. Towards a unified analysis of kernel-based methods under covariate shift. *Advances in Neural Information Processing Systems*, 36:73839–73851, 2023.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *The Twelfth International Conference on Learning Representations*, 2024.

Magnus R. Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*, 49(6):409–436, 1952. doi: 10.6028/jres.049.044.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2 edition, 2021.

Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11):3735–3745, 2009. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2009.04.009.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR, 2021.

Samory Kpotufe and Guillaume Martinet. Marginal singularity and the benefits of labels in covariate-shift. *The Annals of Statistics*, 49(6):3299–3323, 2021.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML 2013 Workshop on Challenges in Representation Learning*, 2013.

Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *Advances in Neural Information Processing Systems*, 34:22968–22981, 2021.

Molei Liu, Yi Zhang, Katherine P. Liao, and Tianxi Cai. Augmented transfer regression learning with semi-non-parametric nuisance models. *Journal of Machine Learning Research*, 24(293):1–50, 2023.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least-squares: Fast rates for regularized empirical risk minimization through self-concordance. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2294–2340. PMLR, 2019. doi: 10.48550/arxiv.1902.03046.

Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition, 1989.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238, 23972327.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. The MIT Press, 2008. ISBN 9780262255103. doi: 10.7551/mitpress/9780262170055.001.0001.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000. ISSN 0378-3758. doi: https://doi.org/10.1016/S0378-3758(00)00115-4.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., 2020.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul Von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, December 2008. ISSN 0020-3157, 1572-9052. doi: 10.1007/s10463-008-0197-x.

Rui Tuo and Lu Zou. Asymptotic theory for linear functionals of kernel ridge regression. Technical Report SS-2024-0256, Statistica Sinica (Preprint), 2024.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer: New York, 1999.

M.J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2019.

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *The Annals of Statistics*, 54(1):252–276, 2026.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations (ICLR)*, 2021.

Appendices

A Preliminaries **15**

A.1 Notations 15

A.2 Useful observations on the GLM excess risk 16

A.3 Kernel GLM examples and assumptions 16

A.4 Proof structure 17

B Proof sketch and key intermediate results **17**

B.1 Generic model selection analysis with GLM loss 17

B.2 Analysis of the KRGLM estimator 19

B.3 Putting things together 21

C Proof of Theorem 4.1 and corollaries **22**

C.1 Fixed-design analysis 22

C.2 Analysis of the selected candidate 23

C.3 Proof of the main Theorem 4.1 25

C.4 Proof of Corollary 4.1 26

C.5 The implicit requirement on the number of target covariates 27

C.6 Optimal $L^2(\mathcal{X};\mathcal{P})$ prediction rate 27

D Proofs of Appendix C **28**

D.1 Proof of Lemma C.1 28

D.2 Proof of Lemma C.2 30

D.3 Proof of Lemma C.3 30

D.4 Proof of Lemma C.4 30

D.5 Proof of Lemma C.5 31

D.6 Proof of Corollary C.1 32

E Proofs of Appendix B.2's pointwise convergence rate **33**

E.1 Coarse-grained analysis of the estimator consistency 34

E.2 Proof of Theorem B.2 37

F Proof of Appendix B.1's in-sample oracle inequality **42**

G Useful lemmas **46**

G.1 Lemmas for the GLM estimator 46

G.2 Technical lemmas 47

G.3 Relating the source covariance operator to the target's via the effective sample size 48

H Complement to the numerical experiments from Section 5 and ablation study **49**

H.1 Implementation details 49

H.2 Additional details and results for synthetic data experiment 49
H.3 Ablation study: validating the undersmoothed imputation penalty 50
H.4 Details about the real data experiments 51
H.5 WILDS benchmark evaluation 52

Notations The constants $c_1, c_2, C_1, C_2, \cdots$ may differ from line to line. We use the symbol $[n]$ as a shorthand for $\{1, 2, \cdots, n\}$ and $|\cdot|$ to denote the absolute value of a real number or cardinality of a set. For nonnegative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ or $a_n = O(b_n)$ if there exists a positive constant C such that $a_n \leq C b_n$. In addition, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n; a_n = o(b_n)$ if $a_n = O(c_n b_n)$ for some $c_n \to 0$. Notations with tildes (e.g., $\widetilde{O}$) hide logarithmic factors. For a matrix $\boldsymbol{A}$, we use $\|\boldsymbol{A}\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|\boldsymbol{A}\boldsymbol{x}\|_2$ to denote its spectral norm. For a bounded linear operator $\boldsymbol{A}$ between two Hilbert spaces $\mathbb{H}_1$ and $\mathbb{H}_2$, we define its operator norm $\|\boldsymbol{A}\| = \sup_{\|\boldsymbol{u}\|_{\mathbb{H}_1}=1} \|\boldsymbol{A}\boldsymbol{u}\|_{\mathbb{H}_2}$. For a bounded, self-adjoint linear operator $\boldsymbol{A}$ mapping a Hilbert space $\mathbb{H}$ to itself, we write $\boldsymbol{A} \succeq 0$ if $\langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{u} \rangle \geq 0$ holds for all $\boldsymbol{u} \in \mathbb{H}$; in that case, we define $\|\boldsymbol{u}\|_{\boldsymbol{A}} = \sqrt{\langle \boldsymbol{u}, \boldsymbol{A}\boldsymbol{u} \rangle}$. For any $\boldsymbol{u}$ and $\boldsymbol{v}$ in a Hilbert space $\mathbb{H}$, their tensor product $\boldsymbol{u} \otimes \boldsymbol{v}$ is a linear operator from $\mathbb{H}$ to itself that maps any $\boldsymbol{w} \in \mathbb{H}$ to $\langle \boldsymbol{v}, \boldsymbol{w} \rangle \boldsymbol{u}$. Define $\|X\|_{\psi_\alpha} = \sup_{p \geq 1} \left\{ p^{-1/\alpha} \mathbb{E}^{1/p} |X|^p \right\}$ for a random variable X and $\alpha \in \{1, 2\}$. If $\boldsymbol{X}$ is a random element in a separable Hilbert space $\mathbb{H}$, then we let $\|\boldsymbol{X}\|_{\psi_\alpha} = \sup_{\|\boldsymbol{u}\|_{\mathbb{H}}=1} \|\langle \boldsymbol{u}, \boldsymbol{X} \rangle\|_{\psi_\alpha}$, $\alpha \in \{1, 2\}$.

A Preliminaries

A.1 Notations

To facilitate the discussion and analysis, we introduce some additional notations. When $\mathcal{X} = \mathbb{R}^d$ for some $d < \infty$ and $K(\boldsymbol{z}, \boldsymbol{w}) = \boldsymbol{z}^\top \boldsymbol{w}, \Phi$ is the identity mapping. We can construct design matrices $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times d}, \boldsymbol{X}_0 = (\boldsymbol{x}_{01}, \cdots, \boldsymbol{x}_{0n_0})^\top \in \mathbb{R}^{n_0 \times d}$ and the response vector $\boldsymbol{y} = (y_1, \cdots, y_n)^\top \in \mathbb{R}^n$. Define the index set $\mathcal{T} = \{i \in [n] : (\boldsymbol{x}_i, y_i) \in \mathcal{D}_1\}$ of the data for training candidate models. In addition, let $n_1 = |\mathcal{T}|$ and $n_2 = n - n_1$. Denote by $\boldsymbol{X}_1 \in \mathbb{R}^{|\mathcal{T}| \times d}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{(n-|\mathcal{T}|) \times d}$ the sub-matrices of $\boldsymbol{X}$ by selecting rows whose indices belong to $\mathcal{T}$ and $[n] \backslash \mathcal{T}$, respectively. Similarly, denote by $\boldsymbol{y}_1 \in \mathbb{R}^{|\mathcal{T}|}$ and $\boldsymbol{y}_2 \in \mathbb{R}^{n-|\mathcal{T}|}$ the sub-vectors of $\boldsymbol{y}$ induced by those index sets.

The design matrices can be generalized to operators when $\mathcal{X}$ and K are general. In that case, $\boldsymbol{X}$ and $\boldsymbol{X}_0$ are bounded linear operators defined through

$$\boldsymbol{X} : \mathbb{H} \to \mathbb{R}^n, \boldsymbol{\theta} \mapsto \left(\langle \Phi(\boldsymbol{x}_1), \boldsymbol{\theta} \rangle, \cdots, \langle \Phi(\boldsymbol{x}_n), \boldsymbol{\theta} \rangle \right)^\top,$$
$$\boldsymbol{X}_0 : \mathbb{H} \to \mathbb{R}^{n_0}, \boldsymbol{\theta} \mapsto \left(\langle \Phi(\boldsymbol{x}_{01}), \boldsymbol{\theta} \rangle, \cdots, \langle \Phi(\boldsymbol{x}_{0n_0}), \boldsymbol{\theta} \rangle \right)^\top.$$

We can define $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ similarly. With slight abuse of notation, we use $\boldsymbol{X}^\top$ to refer to the adjoint of $\boldsymbol{X}$, i.e. $\boldsymbol{X}^\top : \mathbb{R}^n \to \mathbb{H}, \boldsymbol{u} \mapsto \sum_{i=1}^n u_i \Phi(\boldsymbol{x}_i)$. Similarly, we can define $\boldsymbol{X}_0^\top, \boldsymbol{X}_1^\top$ and $\boldsymbol{X}_2^\top$. Let $\widehat{\boldsymbol{\Sigma}}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} \Phi(\boldsymbol{x}_{0i}) \otimes \Phi(\boldsymbol{x}_{0i}), \widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{n_1} \sum_{i \in \mathcal{T}} \Phi(\boldsymbol{x}_i) \otimes \Phi(\boldsymbol{x}_i)$ and $\widehat{\boldsymbol{\Sigma}}_2 = \frac{1}{n_2} \sum_{i \in [n] \backslash \mathcal{T}} \Phi(\boldsymbol{x}_i) \otimes \Phi(\boldsymbol{x}_i)$. We have $\widehat{\boldsymbol{\Sigma}}_j = n_j^{-1} \boldsymbol{X}_j^\top \boldsymbol{X}_j$ for all $j \in \{0, 1, 2\}$.

We will also use the following notation for any $\lambda > 0$:

$$\hat{f}_\lambda(\cdot) = \left\langle \Phi(\cdot), \hat{\boldsymbol{\theta}}_\lambda \right\rangle, \quad f^*(\cdot) = \langle \Phi(\cdot), \boldsymbol{\theta}^* \rangle, \quad \tilde{f}(\cdot) = \left\langle \Phi(\cdot), \tilde{\boldsymbol{\theta}} \right\rangle,$$

where $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\tilde{\lambda}}$.

Also denote $g(\boldsymbol{\theta}) := (a'(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) - y_i)_i^\top$ the vector of residuals such that $\boldsymbol{\varepsilon}_i = g_i(\boldsymbol{\theta}^*)$.

Remark A.1 (Convention)**.** We will adopt the convention that the minimum of a function f over an empty set is infinite:

$$\min_{\lambda \in \emptyset} f(\lambda) = +\infty.$$

A.2 Useful observations on the GLM excess risk

Note that, on the domain where Assumption 4.4 holds, such that the second derivative of the log-partition function is bounded, $k \leq a'' \leq K$, then, for the true excess risk, we have by Taylor-Lagrange:

$$\frac{k}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{Q}}\left|f(\boldsymbol{x})-f^*(\boldsymbol{x})\right|^2 \leq \mathcal{R}(f) \leq \frac{K}{2}\mathbb{E}_{\boldsymbol{x}\sim\mathcal{Q}}\left|f(\boldsymbol{x})-f^*(\boldsymbol{x})\right|^2, \tag{A.1}$$

and we recognize the usual mean squared estimation error under the target distribution $\mathcal{Q}$.

Likewise, for the finite-sample version, we get

$$\frac{k}{2n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i)-f^*(\boldsymbol{x}_i)\right)^2 \leq \mathcal{R}_{in}(f) \leq \frac{K}{2n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i)-f^*(\boldsymbol{x}_i)\right)^2. \tag{A.2}$$

Remark A.2 (Excess risk bounds)**.** Based on Eqs. (A.1) and (A.2) and the notations introduced, we have for any $\lambda > 0$:

$$\frac{k}{2}\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\Sigma_0}^2 \leq \mathcal{R}(\hat{f}_\lambda) \leq \frac{K}{2}\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}_0}^2,$$

and

$$\frac{k}{2}\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\hat{\Sigma}_0}^2 \leq \mathcal{R}_{in}(\hat{f}_\lambda) \leq \frac{K}{2}\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\hat{\boldsymbol{\Sigma}}_0}^2.$$

Remark A.3. The excess risk under the GLM loss is a type of Bregman divergence. The Bregman divergence of the strictly convex and continuously differentiable log-partition function a is defined as:

$$D_a(x,y) := a(x) - a(y) - a'(y)(x-y), \forall x, y \in \Omega,$$

where Ω is a closed convex set in $\mathbb{R}$. Hence, we can rewrite the excess risk and in-sample excess risk respectively as:

$$\mathcal{R}(f) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{Q}}\left[D_a(f(\boldsymbol{x}), f^*(\boldsymbol{x}))\right],$$

and

$$\mathcal{R}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n} D_a(f(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)).$$

A.3 Kernel GLM examples and assumptions

Below are some frequently used kernels and associated RKHSs (Wainwright, 2019).

Example A.1 (Linear and affine kernels)**.** Let $\mathcal{X} \subset \mathbb{R}^d$. The linear kernel $K(\boldsymbol{z},\boldsymbol{w}) = \boldsymbol{z}^\top\boldsymbol{w}$ and the affine kernel $K(\boldsymbol{z},\boldsymbol{w}) = 1 + \boldsymbol{z}^\top\boldsymbol{w}$ yield the RKHS $\mathcal{F}$ of all linear and affine functions respectively, of dimension $D = d$ and $D = d+1$ respectively.

Example A.2 (Polynomial kernels)**.** Let $\mathcal{X} \subset \mathbb{R}^d$. The inhomogeneous polynomial kernel of degree $m \geq 2$ is $K(\boldsymbol{z},\boldsymbol{w}) = (1+\boldsymbol{z}^\top\boldsymbol{w})^m$. It yields the RKHS $\mathcal{F}$ of all polynomials of degree m or less, of dimension $D = \binom{d+m}{m}$.

The function spaces in Examples A.1 and A.2 are finite-dimensional. The following example is infinite-dimensional.

Example A.3 (First-order Sobolev kernel)**.** Let $\mathcal{X} = [0,1]$ and $K(\boldsymbol{z},\boldsymbol{w}) = \min\{\boldsymbol{z},\boldsymbol{w}\}$. Then the first-order Sobolev space is

$$\mathcal{F} = \left\{f : [0,1] \to \mathbb{R} \mid f(0) = 0, \int_0^1 \left|f'(x)\right|^2 dx < \infty\right\}.$$

The boundary condition $f(0) = 0$ excludes many natural regression functions, such as $f^*(x) = 1.5\cos(2\pi x)$ in Section 5.1. As with the affine kernel of Example A.1, this restriction is removed by adding a constant to the kernel: $K(\boldsymbol{z},\boldsymbol{w}) = 1 + \min\{\boldsymbol{z},\boldsymbol{w}\}$ generates the RKHS

$$\mathcal{F} = \left\{f : [0,1] \to \mathbb{R} \mid \int_0^1 \left|f'(x)\right|^2 dx < \infty\right\}, \qquad \|f\|_{\mathcal{F}}^2 = f(0)^2 + \int_0^1 \left|f'(x)\right|^2 dx,$$

which contains $f^*(x) = 1.5\cos(2\pi x)$, with $\|f^*\|_{\mathcal{F}} \approx 6.8$. This is the kernel used in our synthetic experiments. Since the added term is a rank-one perturbation of the integral operator, Weyl's inequality preserves the decay $\mu_j \lesssim j^{-2}$, so Assumption 4.1 still holds.

Table 3: Curvature and smoothness constants for the log-partition function a on $[-2C, 2C]$.

Model	$a(z)$	k	K	L
Linear	$\frac{z^2}{2}$	1	1	0
Logistic	$\log(1+e^z)$	$\frac{e^{2C}}{(1+e^{2C})^2}$	$\frac{1}{4}$	$\frac{1}{10}$
Poisson	e^z	e^{-2C}	e^{2C}	e^{2C}

About Assumption 4.1. The polynomial upper-bound on the spectral decay is a smoothness condition for the kernel: a faster decay (larger α) corresponds to a smoother RKHS, as high-frequency eigendirections contribute less to the kernel. It is satisfied by the first-order Sobolev kernel on $\mathcal{X} = [0, 1]$ and its constant-augmented version (Example A.3) with $\alpha = 1$ under $\mathcal{Q} = \mathcal{U}[0, 1]$. It is also trivially satisfied for any α by finite-rank kernels such as linear and polynomial kernels (Examples A.1 and A.2) which have at most D non-zero eigenvalues.

The uniform boundedness of eigenfunctions holds for linear and polynomial kernels on compact domains, as well as for the first-order Sobolev kernel and its constant-augmented version, whose eigenfunctions are sinusoidal.

Under Assumption 4.1, the feature map induced by the kernel K can be written as

$$\Phi : \boldsymbol{x} \mapsto (\sqrt{\mu_1}\phi_1(\boldsymbol{x}),\, \sqrt{\mu_2}\phi_2(\boldsymbol{x}),\, \ldots)\,.$$

These assumptions on the kernel imply that the covariates are bounded almost surely: there exists a constant $M_2 < \infty$ such that $\|\Phi(\boldsymbol{x})\| \leq M_2$ holds almost surely for $\boldsymbol{x}$ from $\mathcal{P}$ and $\mathcal{Q}$. Then, together with Assumption 2.1, $\forall \boldsymbol{x} \in \operatorname{supp}(\mathcal{P}), |f^*(\boldsymbol{x})| \leq C := RM_2$.

About Assumption 4.4. Table 3 provides simple bounds for the corresponding GLM log-partition constants in the three canonical examples.

A.4 Proof structure

Appendix B presents the two building blocks of our proof and explains how to combine them informally: those are the in-sample oracle-inequality (Theorem B.1) proved in Appendix F and the pointwise convergence rate of the KRGLM estimator (Theorem B.2) proved in Appendix E. Appendix C formalizes this reasoning to prove our main Theorem 4.1. To keep the logic of the proof clear, Appendix C states and combines a number of results that are proved in Appendix D. Throughout, we make use of technical lemmas presented in Appendix G.

B Proof sketch and key intermediate results

Let us summarize the proof idea outlined at the end of Section 4.2. Our main result hinges on two ingredients. Appendix B.1 establishes an in-sample oracle inequality for model selection via pseudo-labels under the GLM loss, characterizing how the imperfection of the pseudo-labels propagates into the excess risk of the selected model. This result is generic and independent of the RKHS structure. Appendix B.2 then shows how, in our kernel GLM setting, each term of this oracle inequality — the imputation error decomposition and the candidate excess risk bounds — can be controlled via a Taylor expansion of the regularized empirical loss's gradient around f^*, yielding the pointwise convergence rate of Theorem B.2 as a key intermediate result. Appendix B.3 explains how these two ingredients are combined to yield Theorem 4.1.

B.1 Generic model selection analysis with GLM loss

At the core of our analysis, we need to understand how the pseudo-labeler impacts the model selection. We consider the task of selecting an optimal candidate from a set of m deterministic models $f_1, \ldots, f_m$ to approximate an unknown regression function f^*. In our GLM framework, f^* characterizes the conditional mean of the response through the canonical link $a'(f^*(\boldsymbol{x}))$. We work under a fixed design of unlabeled samples $\{\boldsymbol{x}_i\}_{i=1}^n$, for which pseudo-labels are generated via an imperfect imputation model $\tilde{f}$. By treating the candidate models and design points as deterministic,

and the imputer as the sole source of randomness (reflecting its dependence on external training data), we isolate the effect of pseudo-label errors from the finite-sample fluctuations of the target covariates, which are handled separately as described in Appendix B.3.

We decompose the pseudo-labels as

$$\tilde{f}(\boldsymbol{x}_i) := f^*(\boldsymbol{x}_i) + B_i + \tilde{\varepsilon}_i, \tag{B.1}$$

where $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_i)_i$ is a centered random vector and $\boldsymbol{B} = (B_i)_i$ is a bias term. They represent the random error and bias of the imputation model.

Recall that the in-sample risk in a GLM framework is given by the negative log-likelihood:

$$R_{in}(f) = \frac{1}{n}\sum_{i=1}^{n}(a(f(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i))f(\boldsymbol{x}_i)), \tag{B.2}$$

and the in-sample excess risk is thus given by

$$\mathcal{R}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n}\Big[a(f(\boldsymbol{x}_i)) - a(f^*(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i))[f(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)]\Big]. \tag{B.3}$$

Since the true function f^* is unknown, we will use as in Eq. (3.1) a proxy for the in-sample excess risk, a pseudo excess risk for model selection purposes:

$$\widehat{\mathcal{R}}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n}\Big[a(f(\boldsymbol{x}_i)) - a(\tilde{f}(\boldsymbol{x}_i)) - a'(\tilde{f}(\boldsymbol{x}_i))[f(\boldsymbol{x}_i) - \tilde{f}(\boldsymbol{x}_i)]\Big]. \tag{B.4}$$

The pseudo-labeling strategy is thus to choose

$$\hat{j} = \operatorname*{argmin}_{j\in[m]} \hat{\mathcal{R}}_{in}(f_j), \tag{B.5}$$

and we denote the selected model by $\hat{f} = f_{\hat{j}}$.

We will only impose some regularity conditions, that are verified by the covariate shift setting of the previous section, as shown in the appendix.

Assumption B.1. *We make the following assumptions:*

1. *The image of the covariates under the true model is bounded almost surely, i.e. there exists a constant $C > 0$ such that $|f^*(\boldsymbol{x}_i)| \leq C$, $\forall i$.*
2. *a is twice differentiable, locally strongly convex and a'' is locally upper bounded and Lipschitz, i.e. there exist finite constants $k > 0, K > 0, L > 0$ such that $\forall x, y \in [-2C, 2C]$, $K \geq a''(x) \geq k$ and $|a''(x) - a''(y)| \leq L|x - y|$.*
3. *With probability at least $1-\delta/2$, both $\max_{i\in[n]} |\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)| \leq \eta$ and $\|\boldsymbol{B}\|_2 \leq b$, for constants $b \geq 0$ and $\eta < \min\{k/L, C, \sqrt{kK}/L\}$.*
4. *The imputation random error $\tilde{\boldsymbol{\varepsilon}} = (\tilde{\varepsilon}_1, \ldots, \tilde{\varepsilon}_n)^\top$ is a zero-mean sub-exponential random vector: $\exists \tilde{\sigma} > 0$, $\|\tilde{\boldsymbol{\varepsilon}}\|_{\psi_1} \leq \tilde{\sigma}$.*

Now the objective is to bound the excess risk of the selected model $\hat{f}$, relative to the excess risk of the model selected by an oracle that knows the true function f^*. We denote by f_{j^*} this oracle model, with

$$j^* = \operatorname*{argmin}_{j\in[m]} \mathcal{R}_{in}(f_j).$$

The following theorem is an oracle inequality for the in-sample excess risk of the selected model:

Theorem B.1 (Oracle inequality)**.** *Let $\delta \in (0,1)$. Then there exists a universal constant $C_1 > 0$ such that under Assumption B.1, the following holds with probability at least $1-\delta$:*

$$\mathcal{R}_{in}(\hat{f}) \leq \inf_{\epsilon>0} \left\{ (1+\epsilon)\mathcal{R}_{in}(f_{j^*}) + \frac{C_1 K}{nk}\left(1+\frac{1}{\epsilon}\right)\xi(\tilde{\sigma}, b, \eta, \delta) \right\}, \tag{B.6}$$

with

$$\xi(\tilde{\sigma}, b, \eta, \delta) = K\tilde{\sigma}^2 \log^2(\frac{m}{\delta}) + Kb^2 + \frac{L^2 n \eta^4}{k}.$$

The proof is deferred to Appendix F. It proceeds by applying the law of cosines for Bregman divergences, which decomposes the true excess risk of any candidate into its pseudo excess risk plus a cross term involving the imputation error $\tilde{f} - f^*$, controlled by the local curvature of a from Assumption B.1. The minimality of the selected model's pseudo risk then transfers to the true risk up to this cross term — a strategy that mirrors the square loss case, but where the Bregman structure of the GLM loss requires additional care to obtain a sharp bound.

The additive overhead in the oracle inequality (B.6) corresponds to the error of the pseudo-labels. It comprises a bias term proportional to b^2/n, where b bounds $\|\boldsymbol{B}\|_2$, a variance term proportional to $\tilde{\sigma}^2 \log^2(m)/n$, and a perturbation term proportional to $L^2\eta^4$, which arises from the nonlinearity of the GLM loss and is shown to be negligible compared to the other two terms in Appendix B.3. Note that in the case of the square loss, $L = 0$, so the perturbation term vanishes and the third item in Assumption B.1 is no longer necessary. The key asymmetry between the bias and variance terms is that $\tilde{\sigma}$ is a scalar sub-exponential proxy for the noise of the imputer, whereas $\|\boldsymbol{B}\|^2 = \sum_{i=1}^n B_i^2$ accumulates the squared pointwise biases over all n design points. Consequently, as long as $\log^2(m)/n$ is small, even a moderate pointwise bias of the imputation model can dominate the variance contribution, making bias the critical quantity to control.

This settles the bias-variance trade-off for the imputation model: we should prioritize an *under-smoothed* imputation model, i.e. with small bias, in order for the pseudo-labels to select a more faithful model. In the case of kernel GLM, the imputation model should thus be trained with a small regularization parameter.

Remark B.1 (Random bias)**.** Assumption B.1 only requires a high-probability bound on $\|\boldsymbol{B}\|_2$: the bias may be random and correlated with $\tilde{\boldsymbol{\varepsilon}}$. This is needed in our application. For KRR, the estimator is linear in the responses and the bias of the imputer is deterministic under fixed design. For kernel GLMs, the estimator has no closed form, so its error is centered around a deterministic surrogate rather than its expectation, and the resulting bias term depends on the noise (Appendix B.2). This is harmless, because $\boldsymbol{B}$ only enters the proof of Theorem B.1 through inequalities that hold for every realization, while concentration is only used for $\tilde{\boldsymbol{\varepsilon}}$ (see Remark F.2).

B.2 Analysis of the KRGLM estimator

We now show how the oracle inequality of Theorem B.1 can be applied in our kernel GLM setting by bounding each of its terms — the imputation error decomposition and the candidate excess risk bounds — via a Taylor expansion of the regularized empirical loss's gradient.

For any positive λ, the estimator $\hat{f}_\lambda$ of Eq. (2.7) is only defined implicitly as the minimizer of a nonlinear objective, unlike the closed-form solution available in kernel ridge regression. For ease of presentation, we derive the following expansion in the finite-dimensional setting with $\mathcal{X} \subset \mathbb{R}^d$ and the linear kernel, writing $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times d}$ for the source design matrix, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ for the noise vector, and $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{x}$. The KRGLM minimization problem Eq. (2.7) becomes

$$\hat{\boldsymbol{\theta}}_\lambda \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ \frac{1}{n}\sum_{i=1}^n \left(a(\boldsymbol{x}_i^\top \boldsymbol{\theta}) - y_i \boldsymbol{x}_i^\top \boldsymbol{\theta} \right) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 \right\}. \tag{B.7}$$

Setting the gradient of the associated loss to zero at $\hat{\boldsymbol{\theta}}_\lambda$ and Taylor-expanding around $\boldsymbol{\theta}^*$ yields:

$$\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = -\boldsymbol{H}^{-1}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda \boldsymbol{\theta}^* \right), \tag{B.8}$$

where the data-dependent Hessian is

$$\boldsymbol{H} = \frac{1}{n}\sum_{i=1}^{n} D_{ii}\,\boldsymbol{x}_i\boldsymbol{x}_i^\top + \lambda \boldsymbol{I}, \qquad D_{ii} = \int_0^1 a''(\boldsymbol{x}_i^\top(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)))\,dt. \tag{B.9}$$

To isolate the dependence on $\hat{\boldsymbol{\theta}}_\lambda$, we introduce the deterministic Hessian $\bar{\boldsymbol{H}} = \frac{1}{n}\sum_{i=1}^{n} a''(\boldsymbol{x}_i^\top\boldsymbol{\theta}^*)\,\boldsymbol{x}_i\boldsymbol{x}_i^\top + \lambda\boldsymbol{I}$, obtained by evaluating a'' at $\boldsymbol{\theta}^*$ rather than along the path. Splitting (B.8) accordingly gives the decomposition:

$$\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = -\Bigg[\underbrace{\lambda\boldsymbol{H}^{-1}\boldsymbol{\theta}^*}_{\text{bias}} + \underbrace{\frac{1}{n}\bar{\boldsymbol{H}}^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}_{\text{variance}} + \underbrace{\frac{1}{n}(\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1})\boldsymbol{X}^\top\boldsymbol{\varepsilon}}_{\text{perturbation}}\Bigg]. \tag{B.10}$$

An equivalent decomposition holds in the general setting via the isometry between the RKHS $\mathcal{F}$ and the feature space $\mathbb{H}$ Eq. (2.4), with $\boldsymbol{X}$ interpreted as a bounded linear operator; see Appendix A.1.

Recall that our Algorithm 1 trains a grid $\{\hat{f}_\lambda\}_{\lambda\in\Lambda}$ of candidate models on the first half $\mathcal{D}_1$ of the source data and an imputation model $\tilde{f} = \hat{f}_{\tilde{\lambda}}$ on the second half $\mathcal{D}_2$, which generates pseudo-labels $\tilde{f}(\boldsymbol{x}_{0i})$ for each target covariate in $\mathcal{D}_0$. To bound the terms entering the oracle inequality of Theorem B.1 applied to $\mathcal{D}_0$, we project decomposition Eq. (B.10) onto the target second-moment operator $\boldsymbol{\Sigma}_0$ for the candidate excess risk $R(\hat{f}_\lambda) - R(f^*)$, and onto the target design points $\{\boldsymbol{x}_{0i}\}$ for the imputation error $\{\tilde{f}(\boldsymbol{x}_{0i}) - f^*(\boldsymbol{x}_{0i})\}_{i=1}^{n_0}$. The main difficulty is analyzing the perturbation term, since $\boldsymbol{H}$ depends on $\hat{\boldsymbol{\theta}}_\lambda$ itself and is correlated with the noise. We resolve this in two steps. First, a localization argument (Lemma E.1) establishes, using only the global strict convexity of a, that $\hat{\boldsymbol{\theta}}_\lambda$ lies in a neighborhood of $\boldsymbol{\theta}^*$ with high probability. Once in this neighborhood, the local properties of a'' from Assumption 4.4 — strong convexity, boundedness, and Lipschitz continuity — all become available. In particular, the Lipschitz continuity ensures that the relative operator error between $\boldsymbol{H}$ and $\bar{\boldsymbol{H}}$ is controlled by the pointwise deviations $|\hat{f}_\lambda(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)|$, reducing the perturbation analysis to a pointwise consistency problem. The following theorem provides the required bound.

Theorem B.2 (Pointwise convergence rate)**.** *Consider a ridge-regularized kernel GLM estimator $\hat{f}_\lambda$ trained on n i.i.d. samples from the source distribution $\mathcal{P}$ via Eq.* (2.7)*. Let $\beta \in (0, 8)$, and choose $\delta \in (0, 1/e]$. Then, under Assumptions 2.1 and 4.1 to 4.4, there exist constants $c, c', c_1, c_2 > 0$, such that for any given $\boldsymbol{x}_0 \in \mathrm{supp}(\mathcal{P})$, when:*

$$\begin{cases} \lambda \le \frac{c_1}{R^4 \log^\beta(n)} \\ \lambda \ge \frac{c_2 M_1^2 \sigma^2 \log(\frac{8}{\delta})(\log n)^{2+\beta/2}}{n} \end{cases},$$

the following inequality holds with probability at least $1-\delta$:

$$\left|\hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0)\right| \le cM_1\left(\sigma\log(\frac{n}{\delta})\frac{\lambda^{-1/4}}{\sqrt{n}} + \lambda^{1/4}\left\|f^*\right\|_{\mathcal{F}}\right). \tag{B.11}$$

When setting λ to its lower bound, the result becomes

$$\left|\hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0)\right| \le c'\sqrt{M_1^3\sigma}\,n^{-1/4}\left(\log(\frac{n}{\delta})\right)^3. \tag{B.12}$$

The proof is deferred to Appendix E.

Bounding the oracle inequality terms via the pointwise rate. We can now carry out the program described above. Projecting (B.10) and applying Theorem B.2 to control the perturbation yields:

1. **Candidate excess risk bounds.** Contracting (B.10) against the target second-moment operator $\boldsymbol{\Sigma}_0$ yields a bound on the excess risk $R(\hat{f}_\lambda) - R(f^*)$ for each candidate. Informally, after applying union bounds, we obtain up to constant and $\log$ factors:

$$\min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} \lesssim \lambda R^2\left\|\boldsymbol{S}_\lambda\right\| + \frac{\sigma^2}{n_1}\mathrm{Tr}(\boldsymbol{S}_\lambda), \tag{B.13}$$

where $\boldsymbol{S}_\lambda = (k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}_0(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2}$. A formal statement can be found in Lemma C.3. Using the effective sample size to relate $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}_0$, one can easily verify that $\mathrm{Tr}(\boldsymbol{S}_\lambda) \leq \frac{2n}{n_{\text{eff}}} \mathrm{Tr}((k\boldsymbol{\Sigma}_0 + \lambda \frac{n}{n_{\text{eff}}}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0)$ and $\|\boldsymbol{S}_\lambda\| \leq \frac{n}{kn_{\text{eff}}}$, as shown in Appendix G.3. Thus taking $n_1 = n/2$, it reduces to

$$\min_{\lambda \in \Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} \lesssim \frac{n}{n_{\text{eff}}}\lambda R^2 + \frac{\sigma^2}{n_{\text{eff}}} \mathrm{Tr}((k\boldsymbol{\Sigma}_0 + \frac{n}{n_{\text{eff}}}\lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0), \tag{B.14}$$

as proved in Appendix C.3.

2. **Imputation error decomposition.** Following the insight of Appendix B.1, our Algorithm 1 trains the imputation model with a small bias with respect to our Λ grid, by choosing $\tilde{\lambda} \asymp n^{-1}$, which also corresponds to the optimal pointwise convergence rate as per Theorem B.2. Applying (B.10) to $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\tilde{\lambda}}$ and projecting onto the target covariates $\{\boldsymbol{x}_{0i}\}_{i=1}^{n_0}$ yields a decomposition of $\{\tilde{f}(\boldsymbol{x}_{0i}) - f^*(\boldsymbol{x}_{0i})\}_{i=1}^{n_0}$ into the bias, variance, and perturbation terms entering $\xi(\tilde{\sigma}, b, \eta, \delta)$ in Theorem B.1, and allows us to bound them in Lemma C.1. Unlike in KRR, the resulting bias term is random, which Assumption B.1 allows (Remark B.1). Informally, for this choice of $\tilde{\lambda}$, we find that the squared variance proxy and the square norm of the bias are of the same order:

$$\tilde{\sigma}^2 \lesssim \frac{n_0}{n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|, \quad \|\boldsymbol{B}\|^2 \lesssim \frac{n_0}{n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|, \tag{B.15}$$

where $\hat{\boldsymbol{S}}_{2,\tilde{\lambda}} = (k\hat{\boldsymbol{\Sigma}}_2 + \tilde{\lambda}\boldsymbol{I})^{-1/2}\hat{\boldsymbol{\Sigma}}_0(k\hat{\boldsymbol{\Sigma}}_2 + \tilde{\lambda}\boldsymbol{I})^{-1/2}$. To bound the perturbation term $n_0\eta^4$, we recall that η bounds $\max_{i\in[n_0]}\left|\tilde{f}(\boldsymbol{x}_{0i}) - f^*(\boldsymbol{x}_{0i})\right|$ with high probability. Substituting the rate $\eta \lesssim n_2^{-1/4}$ from Theorem B.2 yields

$$n_0\eta^4 \lesssim \frac{n_0}{n_2}. \tag{B.16}$$

Pointwise rate. We conclude this section by commenting on the rate obtained in Theorem B.2. We obtain a $O(n^{-1/4})$ pointwise convergence rate of the ridge-regularized kernel GLM estimator when $\lambda \asymp n^{-1}$. This behavior is consistent with the sharp asymptotic theory for linear functionals of kernel ridge regression developed by Tuo and Zou (2024), who analyze point evaluation in Sobolev-equivalent RKHSs: when the input space $\mathcal{X}$ is a convex compact subset of $\mathbb{R}^d$ and $\mathcal{F}$ is equivalent to a Sobolev space of order $m > d/2$, they show in Theorem 7 that $\lambda \asymp n^{-1}$ yields a rate of $n^{-1/2+d/4m}$ for linear functionals, with matching upper and lower bounds (hence optimality in their framework). In particular, the first-order Sobolev (Brownian) kernel $K(u, v) = \min\{u, v\}$ on $[0, 1]$ ($d = m = 1$) satisfies the quadratic eigendecay $\mu_j \lesssim j^{-2}$ and has uniformly bounded eigenfunctions, so both their results and ours predict the same $O(n^{-1/4})$ pointwise rate.

We also emphasize that this pointwise rate is slower than the $L^2(\mathcal{X}; \mathcal{P})$-prediction rate, reflecting the additional difficulty of controlling a point evaluation functional compared to an integrated L^2 loss. We show in Appendix C.6 how to recover the optimal prediction rate $\left\|\hat{f}_\lambda - f^*\right\|_{L^2}^2 = O(n^{-1/3})$ in our setting, obtained for $\lambda \asymp n^{-2/3}$. This in turn matches the fast rate of $O(n^{-\frac{2r\alpha+\alpha}{2r\alpha+\alpha+1}})$ obtained by Marteau-Ferey et al. (2019) with capacity condition $\alpha = 2$ (corresponding to our eigendecay $\alpha = 1$) and source condition $r = 0$ (i.e. without any additional smoothness assumption on f^* beyond membership in the RKHS) in their Corollary 9. The penalty $\lambda \asymp n^{-2/3}$ lies inside the admissible window of Theorem B.2, where it yields the weaker pointwise rate $O(n^{-1/6})$: the two results optimize different functionals, hence different optimal penalties.

Remark B.2 (Role of β)**.** The parameter $\beta \in (0, 8)$ in Theorem B.2 only trades off the two ends of the admissible window for λ, which is non-empty for n large enough, and affects constants and logarithmic factors only. Since $2+\beta/2 < 6$, the penalty $\tilde{\lambda} = \mu^2 \log^7(n) \log(n_0/\delta)/n$ of Theorem 4.1 satisfies the lower bound for every $\beta \in (0, 8)$ simultaneously. At the lower end of the window, (B.11) gives a logarithmic factor of exponent at most $1 + (2 + \beta/2)/4 \leq 5/2$, reported uniformly as $(\log(n/\delta))^3$ in (B.12), which is why β does not appear there.

B.3 Putting things together

We now combine the two previous ingredients. First, we can substitute the bounds on the imputation error decomposition from Eqs. (B.15) and (B.16) with $n_2 = n/2$ in the oracle inequality of

Theorem B.1. We obtain informally and up to $\log$ factors:

$$R_{in}(\hat{f}) - R_{in}(f^*) \lesssim \min_{\lambda\in\Lambda}\{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)\} + \frac{1}{n}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|.$$

A formal statement can be found in Lemma C.4. By further bridging the empirical second-moment operator $\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}$ to its population counterparts $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_0$ through n_{eff}, and relating the in-sample and out-of-sample excess risks via concentration inequalities, it implies:

$$R(\hat{f}) - R(f^*) \lesssim \min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} + O\left(\frac{1}{n_{\text{eff}}} + \frac{1}{n_0}\right).$$

A formal statement is presented in Corollary C.2. To obtain Theorem 4.1, it then suffices to substitute the candidate excess risk bounds from Eq. (B.14) into the first term. As seen in Corollary 4.1, the overhead due to pseudo-label errors, of order $O(1/n_{\text{eff}} + 1/n_0)$, is negligible compared to the resulting candidate excess risks of order $O(n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}})$, determined by the complexity of the target distribution.

C Proof of Theorem 4.1 and corollaries

In this section, we will use the following notations for $j \in \{1, 2\}$:

$$\hat{\boldsymbol{C}}_{j,\lambda} = (k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}_j\,, \quad \hat{\boldsymbol{S}}_{j,\lambda} = (k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I})^{-1/2}\hat{\boldsymbol{\Sigma}}_0(k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I})^{-1/2},$$

their population version

$$\boldsymbol{C}_\lambda = (k\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}\,, \quad \boldsymbol{S}_\lambda = (k\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}_0(k\boldsymbol{\Sigma} + \lambda\boldsymbol{I})^{-1/2},$$

as well as the hybrid

$$\hat{\boldsymbol{U}}_{j,\lambda} = (k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I})^{-1/2}\boldsymbol{\Sigma}_0(k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I})^{-1/2}.$$

To be able to use the in-sample oracle inequality, we first need to conduct a fixed-design analysis by treating the covariates and the data split as deterministic.

C.1 Fixed-design analysis

Lemma C.1 (Fixed-design KRGLM)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold except that all covariates are deterministic and the data split is fixed. Choose any λ satisfying Theorem B.2 conditions:*

$$\begin{cases} \lambda \le \frac{g(n_1)}{R^4} = \frac{1}{R^4\log^\beta(n_1)} \\ \lambda \ge \left(\frac{M_1\sigma\sqrt{\log(16/\delta)\log n_1}}{\sqrt{n_1}}\right)^2 \frac{1}{g(n_1)^{1/2}} = \frac{M_1^2\sigma^2\log(\frac{16}{\delta})(\log n_1)^{2+\beta/2}}{n_1} \end{cases}$$

Choose any positive semi-definite trace-class linear operator $\boldsymbol{Q} : \mathbb{H} \to \mathbb{H}$ and define $\hat{\boldsymbol{T}}_{1,\lambda} = (k\hat{\boldsymbol{\Sigma}}_1 + \lambda\boldsymbol{I})^{-1/2}\boldsymbol{Q}(k\hat{\boldsymbol{\Sigma}}_1 + \lambda\boldsymbol{I})^{-1/2}$. Then, learning with the first half of the source dataset $\mathcal{D}_1$, there is a universal constant C such that for all $\delta \in (0, 1/e]$, with probability at least $1 - \delta$:

$$\left\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\|_{\boldsymbol{Q}}^2 \le 4\left(\lambda R^2\left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\| + \frac{C\sigma^2\log(1/\delta)\log^2(n_1)}{kn_1}\left(L^2R^2M_1^4\left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\| + \operatorname{Tr}(\hat{\boldsymbol{T}}_{1,\lambda})\right)\right).$$

On the other hand, for $\tilde{\lambda}$ set to its lower bound

$$\frac{M_1^2\sigma^2\log(\frac{16n_0}{\delta})(\log n_2)^{2+\beta/2}}{n_2}.$$

Then, learning on the second half of the source dataset $\mathcal{D}_2$, $\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, where $\hat{\boldsymbol{\theta}}_{\tilde{\lambda}} = \tilde{\boldsymbol{\theta}}$, can be decomposed into three terms $\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{B}_1 + \boldsymbol{B}_2 + \tilde{\boldsymbol{\varepsilon}}$, respectively bias, perturbation and variance, satisfying simultaneously, with probability at least $1 - \delta/2$, for all $\delta \in (0, 1/e]$:

- $\|\boldsymbol{B}_1\|_2^2 \le n_0\tilde{\lambda}R^2\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| \le CM_1^2\sigma^2R^2\frac{n_0}{n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\log(\frac{n_0}{\delta})(\log n_2)^6$,
- $\|\boldsymbol{B}_2\|_2^2 \le CM_1^4\sigma^2\frac{L^2n_0}{k^2n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\log^7(n_2/\delta)$,
- $\|\tilde{\boldsymbol{\varepsilon}}\|_{\psi_1}^2 \le \frac{n_0}{n_2}\frac{\sigma^2}{k}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|$ *(this is deterministic), and* $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2 \le C\frac{n_0}{n_2}\frac{\sigma^2}{k}\operatorname{Tr}(\hat{\boldsymbol{S}}_{2,\tilde{\lambda}})\log^3(n_2/\delta)$.

The variance term is sub-exponential (conditionally on $\boldsymbol{X}_2$). C is universal constant.

The proof is deferred to Appendix D.1. The next lemma allows us to convert the previous empirical error bounds to population guarantees, its proof is in Appendix D.2.

Lemma C.2 (Concentration of second-moment operators)**.** *For any λ satisfying Theorem B.2 conditions when $n_1 = n_2 = n/2$:*

$$\begin{cases}\lambda \le \frac{c_1 g(n)}{R^4} = \frac{c_1}{R^4\log^\beta(n)} \\ \lambda \ge c_2\left(\frac{M_1\sigma\sqrt{\log(32/\delta)}\log n}{\sqrt{n}}\right)^2\frac{1}{g(n)^{1/2}} = \frac{c_2M_1^2\sigma^2\log(\frac{32}{\delta})(\log n)^{2+\beta/2}}{n}\end{cases}$$

, and $\lambda_0 = \frac{16M_2^2\log(28n_0/\delta)}{n_0}$, we have simultaneously with probability at least $1-\delta$:

$$\hat{\boldsymbol{\Sigma}}_0 \preceq \frac{3}{2}\boldsymbol{\Sigma}_0 + \frac{\lambda_0}{2}\boldsymbol{I}\,, \quad k\hat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I} \succeq \frac{1}{2}(k\boldsymbol{\Sigma} + \lambda\boldsymbol{I}), \forall j \in \{1,2\}.$$

Under this event, we have for $j \in \{1,2\}$:

$$\operatorname{Tr}(\hat{\boldsymbol{C}}_{j,\lambda}) \le 2M_1^2\operatorname{Tr}(C_\lambda),$$

and

$$\left\|\hat{\boldsymbol{U}}_{j,\lambda}\right\| \le 2\left\|\boldsymbol{S}_\lambda\right\|\,, \quad \operatorname{Tr}(\hat{\boldsymbol{U}}_{j,\lambda}) \le 2\operatorname{Tr}(\boldsymbol{S}_\lambda),$$

and

$$\left\|\hat{\boldsymbol{S}}_{j,\lambda}\right\| \le 3\left\|\boldsymbol{S}_\lambda\right\| + \frac{\lambda_0}{\lambda}.$$

C.2 Analysis of the selected candidate

From the two previous lemmas, we obtain an excess risk bound for the best candidate in $\{\hat{f}_\lambda\}_{\lambda\in\Lambda}$. The proof can be found in Appendix D.3.

Lemma C.3 (Excess risk of the best candidate)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold. Choose any $\delta \in (0, 1/e]$, and define for any $\lambda > 0$*

$$\mathcal{E}(\lambda,\delta) = K\left(\lambda R^2\left\|\boldsymbol{S}_\lambda\right\| + \frac{\sigma^2\log(m/\delta)\log^2(n_1)}{kn_1}\left(\operatorname{Tr}(\boldsymbol{S}_\lambda) + L^2R^2M_1^4\left\|\boldsymbol{S}_\lambda\right\|\right)\right).$$

Then there exists a universal constant C such that with probability at least $1-\delta$,

$$\min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} \le C\min_{\lambda\in\Lambda\cap J(\delta)}\mathcal{E}(\lambda,\delta),$$

where

$$J(\delta) = \left[\frac{M_1^2\sigma^2\log(1/\delta)(\log n)^{2+\beta/2}}{n}\,, \frac{1}{R^4\log^\beta(n)}\right],$$

with the convention $\min\emptyset = \infty$.

The general oracle inequality from Theorem B.1 combined with Lemma C.1 yields the following result for the in-sample risk, and the proof is in Appendix D.4.

Lemma C.4 (Oracle inequality for in-sample risk)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold, except that all covariates are deterministic and the data split is fixed. Choose any $\delta \in (0, 1/e]$, and let $\tilde{\lambda} = \frac{M_1^2\sigma^2\log(16n_0/\delta)(\log n_2)^{2+\beta/2}}{n_2}$, then define*

$$\hat{\xi}(\tilde{\lambda},\delta) = \log^2(\frac{m}{\delta})\frac{(1+L^2)K^2R^2\sigma^2M_1^6\log^7(n_2/\delta)\log(n_0)}{k^2n_2}\left(\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| + 1\right).$$

There exists a universal constant C such that with probability at least $1-\delta$*,*

$$\sqrt{R_{in}(\hat{f})-R_{in}(f^*)} \le \min_{\lambda\in\Lambda}\sqrt{R_{in}(\hat{f}_\lambda)-R_{in}(f^*)}+C\sqrt{\hat{\xi}(\tilde{\lambda},\delta)}.$$

From (Wang, 2026), we apply lemma D.5. The modified proof is in Appendix D.5.

Lemma C.5 (Bridging in-sample and out-of-sample risks)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold, except that the source covariates are deterministic and the data split is fixed. Choose any* $\delta\in(0,1/e]$ *and assume*

$$\Lambda \subseteq \left[\frac{M_1^2\sigma^2\log(8/\delta)(\log n_1)^{2+\beta/2}}{n_1}, \frac{1}{R^4\log^\beta(n_1)}\right].$$

Then there exists a universal constant C *such that with probability at least* $1-\delta$*,*

$$\max_{\lambda\in\Lambda}\left|\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\Sigma_0}-\left\|\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\|_{\hat{\Sigma}_0}\right| \le C\sqrt{M_1\sigma}\sqrt{\frac{\log^2(n_0m/\delta)\log(m/\delta)}{n_0}}.$$

With the same assumptions and definitions, the previous lemmas lead to:

Corollary C.1. *There exists a universal constant C such that with probability at least* $1-2\delta$*,*

$$\sqrt{R(\hat{f})-R(f^*)} \le \frac{K}{k}\min_{\lambda\in\Lambda}\sqrt{R(\hat{f}_\lambda)-R(f^*)}+C\sqrt{\frac{K}{k}}\left(\sqrt{\hat{\xi}(\tilde{\lambda},\delta)}+\sqrt{\frac{M_1\sigma\log^2(n_0m/\delta)\log(m/\delta)}{n_0}}\right).$$

The proof is in Appendix D.6. Further combining, we get:

Corollary C.2 (Oracle inequality for pseudo-labeling)**.** *Let Assumptions 2.1 and 4.1 to 4.4 hold. Assume* $n_1=n_2=n/2$*. When* $\tilde{\lambda}\asymp\frac{M_1^2\sigma^2\log(n_0/\delta)\log(n)^{2+\beta/2}}{n}$ *and*

$$\Lambda \subseteq \left[\frac{M_1^2\sigma^2\log(1/\delta)(\log n)^{2+\beta/2}}{n}, \frac{1}{R^4\log^\beta(n)}\right],$$

with probability at least $1-\delta$*,*

$$\sqrt{R(\hat{f})-R(f^*)} \le \frac{K}{k}\min_{\lambda\in\Lambda}\sqrt{R(\hat{f}_\lambda)-R(f^*)}+\zeta\cdot\sqrt{\frac{(1+L^2)K^3R^2M_1^6\sigma^2}{k^3}(\frac{1}{n_{\mathit{eff}}}+\frac{1}{n_0})},$$

where ζ *is a polylogarithmic factor in* n, n_0, m, δ^{-1}*.*

It implies up to a polylog factor,

$$R(\hat{f})-R(f^*) \lesssim \frac{K^2}{k^2}\min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda)-R(f^*)\}+\frac{(1+L^2)K^3R^2M_1^6\sigma^2}{k^3}(\frac{1}{n_{\text{eff}}}+\frac{1}{n_0}).$$

Plugging the bound on the second term from Lemma C.3 yields the main theorem.

Proof. In this proof, we use $\lesssim$ to hide polylogarithmic factors in (n, n_0, m, δ^{-1}). We relate the in-sample quantities from Corollary C.1 above to their population version via Lemma C.2.

On the one hand, from Corollary C.1, we have

$$\hat{\xi}(\tilde{\lambda},\delta/4)=\log^2(\frac{4m}{\delta})\frac{(1+L^2)K^2R^2\sigma^2M_1^6\log^7(n/\delta)\log(n_0)}{k^2n}\left(\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|+1\right).$$

On the other hand, the concentration bounds from Lemma C.2 give

$$\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| \le 3\left\|\boldsymbol{S}_{\tilde{\lambda}}\right\|+\frac{\lambda_0}{\tilde{\lambda}} \lesssim \frac{n}{n_{\text{eff}}}+\frac{n}{n_0},$$

where $\lambda_0 = \frac{16M_2^2 \log(56n_0/\delta)}{n_0}$, and we used Eq. (G.3) since $\tilde{\lambda} > \mu^2/n$. Putting everything together, since by construction $n_{\text{eff}} \leq n$, we obtain:

$$\hat{\xi}(\lambda, \delta/4) \lesssim \frac{(1+L^2)K^2R^2\sigma^2M_1^6}{k^2}\left(\frac{1}{n_{\text{eff}}} + \frac{1}{n_0}\right).$$

Since the additional term $\frac{M_1\sigma \log^2(n_0 m/\delta)\log(m/\delta)}{n_0}$ is just another polylog divided by the target sample size, we deduce that with probability at least $1-\delta$,

$$\sqrt{R(\hat{f}) - R(f^*)} \leq \frac{K}{k}\min_{\lambda\in\Lambda}\sqrt{R(\hat{f}_\lambda) - R(f^*)} + \zeta \cdot \sqrt{\frac{(1+L^2)K^3R^2M_1^6\sigma^2}{k^3}(\frac{1}{n_{\text{eff}}} + \frac{1}{n_0})}.$$

□

C.3 Proof of the main Theorem 4.1

Combining via union bound Corollary C.2 and Lemma C.3, we get that, for $\Lambda = \{\lambda_j\}_{j=1}^m$ defined in the theorem, the following holds with probability at least $1-\delta$:

$$R(\hat{f}) - R(f^*) \lesssim \frac{K^2}{k^2}\min_{\lambda\in\Lambda\cap J(\delta)}\{\mathcal{E}(\lambda,\delta)\} + \frac{(1+L^2)K^3R^2M_1^6\sigma^2}{k^3}(\frac{1}{n_{\text{eff}}} + \frac{1}{n_0}),$$

and

$$\begin{aligned}\mathcal{E}(\lambda,\delta) &\leq \frac{K}{k}\left(\lambda R^2 r^{-1} + \frac{\sigma^2\log^2(m/\delta)\log^2(n)}{n}(2r^{-1}\operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \lambda r^{-1}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0) + L^2R^2M_1^4 r^{-1}k^{-1})\right) \\ &\lesssim \frac{K\log^2(n)\log^2(m/\delta)}{k}\left(\lambda R^2\frac{n}{n_{\text{eff}}} + \frac{\sigma^2}{n_{\text{eff}}}\operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \lambda r^{-1}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0)\right),\end{aligned}$$

where we used again that $\|\boldsymbol{S}_\lambda\| \leq r^{-1}k^{-1}$ and $\operatorname{Tr}(\boldsymbol{S}_\lambda) \leq 2r^{-1}\operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \lambda r^{-1}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0)$, and that the last perturbation term is negligible compared to the trace term of order $n_{\text{eff}}^{1/2\alpha_0}$.

The last step is to bridge the discrete grid with the continuous interval. We adapt the proof of theorem 4.1 (Wang, 2026), by replacing the decision set Λ by $\tilde{\Lambda} := \Lambda \cap J(\delta) = \{\lambda_1, ..., \lambda_p\}$, with $\lambda_1 \geq \frac{\mu^2}{n}\log(1/\delta)\log(n)^{2+\beta/2}$ and $\lambda_p \leq \frac{1}{R^4\log(n)^\beta}$.

Denote

$$\min_{\lambda\in\tilde{\Lambda}}\{\underbrace{nR^2\lambda}_{A(\lambda)} + \underbrace{\sigma^2\operatorname{Tr}\left[\left(k\boldsymbol{\Sigma}_0 + \mathrm{r}^{-1}\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Sigma}_0\right]}_{B(\lambda)}\}.$$

We will finish the proof by establishing the following connection between the finite grid and the infinite interval:

$$\min_{\lambda\in\tilde{\Lambda}}\{A(\lambda) + B(\lambda)\} \lesssim \inf_{\lambda>0}\{A(\lambda) + B(\lambda)\} + R^2\mu^2. \tag{C.1}$$

The functions A and B are increasing and decreasing, respectively. From the bound on λ_1 we get $A(\lambda_1) \asymp R^2\mu^2$ and

$$\inf_{0<\lambda\leq\lambda_1}\{A(\lambda) + B(\lambda)\} + R^2\mu^2 \geq B(\lambda_1) + A(\lambda_1). \tag{C.2}$$

From $\lambda_p \asymp \frac{\mu^2}{R^4\log^\beta n}$ we get

$$\inf_{\lambda \geq \lambda_p} \{A(\lambda) + B(\lambda)\} \geq A(\lambda_p) = nR^2 \lambda_p \asymp \frac{n\mu^2}{R^2 \log^\beta n}.$$

Meanwhile,

$$B(\lambda_p) = \sigma^2 \operatorname{Tr}\left[\left(k\boldsymbol{\Sigma}_0 + \mathrm{r}^{-1}\lambda_p \boldsymbol{I}\right)^{-1} \boldsymbol{\Sigma}_0\right] \lesssim \lambda_p^{-1/2\alpha_0} \asymp \log^{\beta/2\alpha_0}(n) \lesssim A(\lambda_p),$$

since $A(\lambda_p)$ grows linearly in n up to logarithmic factors.

Therefore,

$$\inf_{\lambda \geq \lambda_p} \{A(\lambda) + B(\lambda)\} \gtrsim A(\lambda_p) + B(\lambda_p). \tag{C.3}$$

Next, we bridge the grid $\tilde{\Lambda}$ and the interval $[\lambda_1, \lambda_p]$. For any $\lambda \in [\lambda_j, \lambda_{j+1}]$, we have $A(\lambda) \geq A(\lambda_j)$ and $k\boldsymbol{\Sigma}_0 + \mathrm{r}^{-1}\lambda \boldsymbol{I} \preceq (\lambda/\lambda_j)\left(k\boldsymbol{\Sigma}_0 + \mathrm{r}^{-1}\lambda_j \boldsymbol{I}\right)$. Hence,

$$A(\lambda) + B(\lambda) \geq A(\lambda_j) + \frac{\lambda_j}{\lambda} B(\lambda_j) \geq \frac{A(\lambda_j) + B(\lambda_j)}{2}.$$

We get

$$\inf_{\lambda_1 \leq \lambda \leq \lambda_p} \{A(\lambda) + B(\lambda)\} \geq \frac{1}{2} \min_{\lambda \in \tilde{\Lambda}} \{A(\lambda) + B(\lambda)\}. \tag{C.4}$$

The claim (C.1) and the theorem follow from Eqs. (C.2) to (C.4).

Note that since $m = \lceil \log_2(n) \rceil + 1$, we make the pre-factor ζ a polylog in (n, n_0, δ^{-1}) only.

C.4 Proof of Corollary 4.1

Proof. Throughout the proof, we use $\lesssim$ to hide polylogarithmic factors and constants independent of n and n_{eff}. From Theorem 4.1, the excess risk is bounded by:

$$R(\hat{f}) - R(f^*) \lesssim \underbrace{\inf_{\rho > 0} \left\{ \rho + \frac{1}{n_{\text{eff}}} \sum_{j=1}^{\infty} \frac{\mu_{0j}}{k\mu_{0j} + \rho} \right\}}_{A(\rho)} + \frac{1}{n_0} + \frac{1}{n_{\text{eff}}}. \tag{C.5}$$

We analyze the term inside the sum. Observe that for any j:

$$\frac{\mu_{0j}}{k\mu_{0j} + \rho} = \frac{1}{k}\left(\frac{k\mu_{0j}}{k\mu_{0j} + \rho}\right) \leq \frac{1}{k} \min\left\{1, \frac{k\mu_{0j}}{\rho}\right\}. \tag{C.6}$$

Substituting this bound and using the decay assumption $\mu_{0j} \leq cj^{-2\alpha}$, we split the sum at an integer index J:

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{\mu_{0j}}{k\mu_{0j} + \rho} &\leq \frac{1}{k} \sum_{j=1}^{J} 1 + \frac{1}{k} \sum_{j=J+1}^{\infty} \frac{kcj^{-2\alpha}}{\rho} \\
&= \frac{J}{k} + \frac{c}{\rho} \sum_{j=J+1}^{\infty} j^{-2\alpha} \\
&\lesssim \frac{J}{k} + \frac{J^{1-2\alpha}}{\rho}.
\end{aligned}$$

Substituting this back into $A(\rho)$:

$$A(\rho) \lesssim \rho + \frac{1}{n_{\text{eff}}}\left(\frac{J}{k} + \frac{J^{1-2\alpha}}{\rho}\right). \tag{C.7}$$

We choose the regularization parameter $\rho = n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}}$ and the cutoff index $J = \lceil \rho^{-\frac{1}{2\alpha}} \rceil \asymp n_{\text{eff}}^{\frac{1}{2\alpha+1}}$. Checking the balance of terms (ignoring k and constants for the rate):

- $\rho = n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}}$.
- $\frac{J}{n_{\text{eff}}} \asymp n_{\text{eff}}^{\frac{1}{2\alpha+1}-1} = n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}}$.
- $\frac{J^{1-2\alpha}}{\rho n_{\text{eff}}} \asymp \frac{(n_{\text{eff}}^{\frac{1}{2\alpha+1}})^{1-2\alpha}}{n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}} n_{\text{eff}}} = n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}}$.

Thus, $A(\rho) \lesssim n_{\text{eff}}^{-\frac{2\alpha}{2\alpha+1}}$, completing the proof. □

C.5 The implicit requirement on the number of target covariates

We detail Remark 4.3. In Corollary 4.1, the pseudo-label term n_0^{-1} is negligible compared with the estimation rate if and only if $n_0 \gtrsim n_{\text{eff}}^{2\alpha/(2\alpha+1)}$. Since $2\alpha/(2\alpha+1) < 1$, this requirement is sublinear in n_{eff}, hence in n; for $\alpha = 1$, it reads $n_0 \gtrsim n_{\text{eff}}^{2/3}$, and the exponent only approaches 1 as $\alpha \to \infty$. As the requirement decreases with n_{eff}, it is easiest to meet under hard shifts: for $\alpha = 1$ and $n_{\text{eff}} \asymp \sqrt{n}$ (the Sobolev example of Section 4.2), it reduces to $n_0 \gtrsim n^{1/3}$. It binds under mild shifts, $n_{\text{eff}} \asymp n$, with scarce target covariates (e.g., a rare subpopulation): if $n_0 \lesssim n_{\text{eff}}^{2\alpha/(2\alpha+1)}$, the bound is driven by the $O(n_0^{-1})$ pseudo-label error regardless of how much source data is available. The bound remains valid in all regimes.

C.6 Optimal $L^2(\mathcal{X}; \mathcal{P})$ prediction rate

We proved in Theorem B.2 a pointwise convergence rate of $O(n^{-1/4})$ for a generic ridge-regularized kernel GLM estimator, with polynomial eigendecay $\mu_j \lesssim j^{-2}$. Our analysis also allows us to derive the optimal L^2 prediction rate of this estimator. We will omit the details of the proof because it follows the exact same steps as our main proof. Below we present a proof sketch. $\lesssim$ hides constants and $\log$ factors.

We can use part 1 of Lemma C.1 with $\boldsymbol{Q} = \boldsymbol{\Sigma}$, combined with the concentration results of Lemma C.2 to obtain an equivalent form of Lemma C.3 for one distribution only. Further using the fact that $\|\boldsymbol{C}_\lambda\| \lesssim 1$ and $\operatorname{Tr}(\boldsymbol{C}_\lambda) \lesssim \lambda^{-1/2}$ (as per Remark G.1), we obtain

$$\left\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}}^2 \lesssim \lambda R^2 + \frac{\sigma^2 \lambda^{-1/2}}{n}.$$

Minimizing on λ yields the announced $\lambda^* \asymp n^{-2/3}$ and $\left\|\hat{\boldsymbol{\theta}}_{\lambda^*} - \boldsymbol{\theta}^*\right\|_{\boldsymbol{\Sigma}}^2 \lesssim n^{-2/3}$.

Since by definition

$$\left\|\hat{f}_\lambda - f^*\right\|_{L^2}^2 = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}}\left[\left\langle \Phi(\boldsymbol{x}), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle^2\right] = \left\langle \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*, \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\rangle,$$

we directly get:

$$\left\|\hat{f}_{\lambda^*} - f^*\right\|_{L^2} \lesssim n^{-1/3},$$

and the GLM excess risk

$$\mathcal{R}(f_{\lambda^*}) \lesssim n^{-2/3}.$$

D Proofs of Appendix C

D.1 Proof of Lemma C.1

The key decomposition to use for this lemma comes from the proof of Theorem B.2 in Appendix E.2. With the same notations [3]:

$$\forall \lambda > 0, -(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) = \lambda \boldsymbol{H}^{-1}\boldsymbol{\theta}^* + (\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1})\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\varepsilon} + \bar{\boldsymbol{H}}^{-1}\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\varepsilon}.$$

The first term is the bias term, the last term is a sub-exponential variance term, and the middle term is a perturbation that we regroup with the bias.

For the first part of the lemma, let $\boldsymbol{Q}$ be any PSD trace-class operator and λ in the specified range, working on the first half of the dataset. Per Theorem B.2 (and intermediate results obtained during its proof), with probability at least $1 - \delta/2$, the following inequalities hold simultaneously:

$$\boldsymbol{H}_1 \succeq k\hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I}, \tag{D.1}$$

$$\boldsymbol{H}_1^{-1} - \bar{\boldsymbol{H}}_1^{-1} = -\bar{\boldsymbol{H}}_1^{-1/2}(\boldsymbol{\Delta}_1(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1})\bar{\boldsymbol{H}}_1^{-1/2}, \tag{D.2}$$

$$\|\boldsymbol{\Delta}_1\| \lesssim \frac{LM_1}{k}\left(\frac{\sigma \log(n_1/\delta)\lambda^{-1/4}}{\sqrt{n_1}} + R\lambda^{1/4}\right), \tag{D.3}$$

$$\left\|\frac{1}{\sqrt{n_1}}\bar{\boldsymbol{H}}_1^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|^2 \leq \sigma^2 \log^2(n_1)\log(8/\delta)\operatorname{Tr}(\hat{\boldsymbol{C}}_{1,\lambda}). \tag{D.4}$$

Then:

$$\left\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\|_Q^2 \leq 4\left[\left\|\lambda \boldsymbol{H}_1^{-1}\boldsymbol{\theta}^*\right\|_Q^2 + \left\|(\boldsymbol{H}_1^{-1} - \bar{\boldsymbol{H}}_1^{-1})\frac{1}{n_1}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|_Q^2 + \left\|\bar{\boldsymbol{H}}_1^{-1}\frac{1}{n_1}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|_Q^2\right].$$

We study the three terms separately.

- For the first term, use Eq. (D.1)

$$\begin{aligned}\left\|\lambda \boldsymbol{H}_1^{-1}\boldsymbol{\theta}^*\right\|_Q^2 &\leq \lambda^2 R^2 \left\|\boldsymbol{Q}^{1/2}\boldsymbol{H}_1^{-1}\right\|^2 \leq \lambda^2 R^2 \left\|\boldsymbol{H}_1^{-1}\right\| \left\|\boldsymbol{Q}^{1/2}\boldsymbol{H}_1^{-1}\boldsymbol{Q}^{1/2}\right\| \\ &\leq \lambda R^2 \left\|\boldsymbol{Q}^{1/2}(k\hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I})^{-1}\boldsymbol{Q}^{1/2}\right\| = \lambda R^2 \left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\|,\end{aligned}$$

  where we wrote $\left\|\boldsymbol{Q}^{1/2}\boldsymbol{H}_1^{-1}\right\|^2 = \left\|(\boldsymbol{Q}^{1/2}\boldsymbol{H}_1^{-1/2})\boldsymbol{H}_1^{-1}(\boldsymbol{H}_1^{-1/2}\boldsymbol{Q}^{1/2})\right\|$, used $\left\|\boldsymbol{H}_1^{-1}\right\| \leq 1/\lambda$, and used $\boldsymbol{H}_1 \succeq k\hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I}$ together with the operator monotonicity of $t \mapsto -t^{-1}$ (note that $t \mapsto t^{-2}$ is not operator monotone, so $\boldsymbol{H}_1^{-2}$ cannot be compared directly with $(k\hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I})^{-2}$). The last equality holds because $\left\|\boldsymbol{A}\boldsymbol{A}^\top\right\| = \left\|\boldsymbol{A}^\top \boldsymbol{A}\right\|$ for $\boldsymbol{A} = \boldsymbol{Q}^{1/2}(k\hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I})^{-1/2}$.
- For the second term, use Eqs. (D.2) and (D.4):

$$\begin{aligned}\left\|(\boldsymbol{H}_1^{-1} - \bar{\boldsymbol{H}}_1^{-1})\frac{1}{n_1}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|_Q^2 &= \frac{1}{n_1^2}\left\|\boldsymbol{Q}^{1/2}\bar{\boldsymbol{H}}_1^{-1/2}(\boldsymbol{\Delta}_1(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1})\bar{\boldsymbol{H}}_1^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|^2 \\ &\leq \frac{1}{n_1}\left\|\boldsymbol{Q}^{1/2}\bar{\boldsymbol{H}}_1^{-1/2}\right\|^2 \cdot \left\|(\boldsymbol{\Delta}_1(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1})\right\|^2 \cdot \left\|\frac{1}{\sqrt{n_1}}\bar{\boldsymbol{H}}_1^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|^2 \\ &\lesssim \frac{1}{n_1}\left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\| \|\boldsymbol{\Delta}_1\|^2 \sigma^2 \log^2(n_1)\log(8/\delta)\operatorname{Tr}(\hat{\boldsymbol{C}}_{1,\lambda})\end{aligned}$$

[3] In what follows, we denote by $\boldsymbol{H}_i$, $\boldsymbol{D}_i$, $\boldsymbol{\Delta}_i$ the corresponding quantities defined in the proof of Theorem B.2 when working with dataset $\mathcal{D}_i$.

$$\lesssim \frac{1}{n_1} \left\| \hat{\boldsymbol{T}}_{1,\lambda} \right\| \|\boldsymbol{\Delta}_1\|^2 \sigma^2 \log^2(n_1) \log(8/\delta) M_1^2 \operatorname{Tr}(\boldsymbol{C}_\lambda),$$

where the last inequality uses $\operatorname{Tr}(\hat{\boldsymbol{C}}_{1,\lambda}) \leq 2M_1^2 \operatorname{Tr}(\boldsymbol{C}_\lambda)$.
Recall that $\operatorname{Tr}(\boldsymbol{C}_\lambda) \lesssim \lambda^{-1/2}$ and from Eq. (D.3), we get:

$$\|\boldsymbol{\Delta}_1\|^2 \operatorname{Tr}(\boldsymbol{C}_\lambda) \lesssim \frac{L^2 M_1^2}{k^2} \left(\frac{\sigma^2 \log^2(n_1/\delta)}{n_1 \lambda} + R^2 \right) \lesssim \frac{L^2 M_1^2}{k^2} \left(\frac{1}{M_1^2 \log^{\beta/2}(n_1/\delta)} + R^2 \right) \lesssim \frac{L^2 M_1^2 R^2}{k^2},$$

so that

$$\left\| (\bar{\boldsymbol{H}}_1^{-1} - \bar{\boldsymbol{H}}_1^{-1}) \frac{1}{n_1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1 \right\|_Q^2 \lesssim \frac{L^2}{k^2 n_1} \left\| \hat{\boldsymbol{T}}_{1,\lambda} \right\| \sigma^2 \log^2(n_1) \log(8/\delta) R^2 M_1^4.$$

- For the third term, use a sub-exponential quadratic form tail bound that holds with probability larger than $1 - \delta/2$:

$$\begin{aligned}
\left\| \bar{\boldsymbol{H}}_1^{-1} \frac{1}{n} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1 \right\|_Q^2 &= \frac{1}{n_1^2} \left\langle \boldsymbol{\varepsilon}_1, \boldsymbol{X}_1 \bar{\boldsymbol{H}}_1^{-1} \boldsymbol{Q} \bar{\boldsymbol{H}}_1^{-1} \boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1 \right\rangle \\
&\leq \frac{\sigma^2}{n_1^2} \operatorname{Tr}(\boldsymbol{X}_1 \bar{\boldsymbol{H}}_1^{-1} \boldsymbol{Q} \bar{\boldsymbol{H}}_1^{-1} \boldsymbol{X}_1^\top) \log^2(n_1) \log(4/\delta) \\
&\leq \frac{\sigma^2}{n_1} \operatorname{Tr}(\bar{\boldsymbol{H}}_1^{-1} \boldsymbol{Q} \bar{\boldsymbol{H}}_1^{-1} \hat{\boldsymbol{\Sigma}}_1) \log^2(n_1) \log(4/\delta) \\
&= \frac{\sigma^2}{n_1} \operatorname{Tr}(\boldsymbol{Q} \bar{\boldsymbol{H}}_1^{-1} \hat{\boldsymbol{\Sigma}}_1 \bar{\boldsymbol{H}}_1^{-1}) \log^2(n_1) \log(4/\delta) \\
&\leq \frac{\sigma^2}{k n_1} \operatorname{Tr}(\boldsymbol{Q} \bar{\boldsymbol{H}}_1^{-1}) \log^2(n_1) \log(4/\delta) \\
&\leq \frac{\sigma^2}{k n_1} \operatorname{Tr}(\boldsymbol{Q} (k \hat{\boldsymbol{\Sigma}}_1 + \lambda \boldsymbol{I})^{-1}) \log^2(n_1) \log(4/\delta) \\
&= \frac{\sigma^2}{k n_1} \operatorname{Tr}(\hat{\boldsymbol{T}}_{1,\lambda}) \log^2(n_1) \log(4/\delta).
\end{aligned}$$

The bound follows by union bound and redefining constant C.

For the second part of the lemma, we work with the second half of the source dataset and $\tilde{\lambda}$ set to the lower bound of the range, so that Theorem B.2 now guarantees a $n^{-1/4}$ consistency rate (up to logarithmic factors), otherwise the three inequalities Eqs. (D.1), (D.2) and (D.4) are still valid with probability at least $1 - \delta/2$ with $\boldsymbol{X}_2$, $\boldsymbol{H}_2$ $\boldsymbol{\varepsilon}_2$... and $\|\boldsymbol{\Delta}_2\|^2 \lesssim \frac{L^2}{k^2} M_1^3 \sigma n_2^{-1/2} \log^6(n_2/\delta)$. We use the same decomposition multiplied by the target design matrix $\boldsymbol{X}_0$:

$$\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{B}_1 + \boldsymbol{B}_2 + \tilde{\boldsymbol{\varepsilon}},$$

where

$$\begin{aligned}
\|\boldsymbol{B}_1\|_2^2 &= \left\| \tilde{\lambda} \boldsymbol{X}_0 \boldsymbol{H}_2^{-1} \boldsymbol{\theta}^* \right\|^2 \\
&\leq \tilde{\lambda}^2 R^2 \left\| \boldsymbol{X}_0 \boldsymbol{H}_2^{-2} \boldsymbol{X}_0^\top \right\| \\
&\leq \tilde{\lambda} R^2 \left\| \boldsymbol{X}_0 \boldsymbol{H}_2^{-1} \boldsymbol{X}_0^\top \right\| \\
&\leq \tilde{\lambda} R^2 \left\| \boldsymbol{X}_0 (k \hat{\boldsymbol{\Sigma}}_2 + \tilde{\lambda} \boldsymbol{I})^{-1} \boldsymbol{X}_0^\top \right\| \\
&= n_0 \tilde{\lambda} R^2 \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\|,
\end{aligned}$$

$$\|\boldsymbol{B}_2\|_2^2 = \left\| \boldsymbol{X}_0 \bar{\boldsymbol{H}}_2^{-1/2} (\boldsymbol{\Delta}_2 (\boldsymbol{I} + \boldsymbol{\Delta}_2)^{-1}) \bar{\boldsymbol{H}}_2^{-1/2} \frac{1}{n_2} \boldsymbol{X}_2^\top \boldsymbol{\varepsilon}_2 \right\|_2^2$$

$$
\begin{aligned}
&\leq \frac{1}{n_2}\left\|\boldsymbol{X}_0\bar{\boldsymbol{H}}_2^{-1/2}\right\|_2^2 \cdot 2\left\|\boldsymbol{\Delta}_2\right\|_2^2 \cdot \left\|\frac{1}{\sqrt{n_2}}\bar{\boldsymbol{H}}_2^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}_2\right\|_2^2 \\
&\lesssim \frac{L^2 n_0}{k^2 n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| \cdot M_1^3\sigma n_2^{-1/2}\log^6(n_2)\cdot\sigma^2\log^2(n_2)\log(8/\delta)\operatorname{Tr}(\hat{\boldsymbol{C}}_{2,\tilde{\lambda}}) \\
&\lesssim \frac{L^2 n_0}{k^2 n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| \cdot M_1^3 n_2^{-1/2}\log^9(n_2/\delta)\cdot\sigma^3 M_1^2\tilde{\lambda}^{-1/2} \quad \text{since } \operatorname{Tr}(\hat{\boldsymbol{C}}_{2,\tilde{\lambda}}) \leq 2M_1^2\operatorname{Tr}(\boldsymbol{C}_\lambda) \lesssim 2M_1^2\tilde{\lambda}^{-1/2} \\
&\lesssim \frac{L^2 n_0}{k^2 n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\| \cdot M_1^3 n_2^{-1/2}\log^9(n_2/\delta)\cdot\sigma^3 M_1^2\cdot n_2^{1/2}M_1^{-1}\sigma^{-1}\log^{-2}(n_2/\delta)\log^{-1}(n_0/\delta) \\
&\lesssim \frac{L^2 n_0}{k^2 n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\log^7(n_2/\delta)\sigma^2 M_1^4,
\end{aligned}
$$

and finally

$$
\begin{aligned}
\|\tilde{\boldsymbol{\varepsilon}}\|_{\psi_1}^2 &= \left\|\frac{1}{n_2}\boldsymbol{X}_0\bar{\boldsymbol{H}}_2^{-1}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}_2\right\|_{\psi_1}^2 \\
&\leq \sigma^2\left\|\frac{1}{n_2}\boldsymbol{X}_0\bar{\boldsymbol{H}}_2^{-1}\boldsymbol{X}_2^\top\right\|_2^2 \\
&= \sigma^2\left\|\boldsymbol{X}_0\bar{\boldsymbol{H}}_2^{-1}\frac{1}{n_2^2}\boldsymbol{X}_2^\top\boldsymbol{X}_2\bar{\boldsymbol{H}}_2^{-1}\boldsymbol{X}_0^\top\right\|_2 \\
&\leq \sigma^2\left\|\boldsymbol{X}_0\bar{\boldsymbol{H}}_2^{-1}\frac{1}{n_2 k}\boldsymbol{X}_0^\top\right\|_2 \quad \text{because} \quad \frac{1}{n_2}\boldsymbol{X}_2^\top\boldsymbol{X}_2 = \hat{\boldsymbol{\Sigma}}_2 \preceq \frac{1}{k}\bar{\boldsymbol{H}}_2 \\
&\leq \frac{\sigma^2 n_0}{k n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|,
\end{aligned}
$$

with the associated tail bound for $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2$, obtained through the quadratic form tail bound of sub-exponential vectors result of Lemma G.2.

D.2 Proof of Lemma C.2

Apply Lemma G.3 with $\gamma = 1/2$ and note that when n is large enough the new lower bound is larger than the one required on the lemma.

D.3 Proof of Lemma C.3

Start by leveraging the observation made in Eq. (A.1): $R(\hat{f}_\lambda) - R(f^*) \leq \frac{K}{2}\left\|\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\|_{\Sigma_0}^2$. Then apply Lemma C.1 with $\boldsymbol{Q} = \boldsymbol{\Sigma}_0$ and union bounds on all candidates to get error bounds with empirical second-moment operators. Further, apply Lemma C.2 to relate them to the population versions. The two first lemmas are only valid in the range of $J(\delta)$. Finally, use the fact that $\min_{\lambda\in\Lambda} \leq \min_{\lambda\in\Lambda\cap J}$.

D.4 Proof of Lemma C.4

We want to apply the in-sample oracle inequality from Theorem B.1 to our context. The samples of interest are the target covariates $\{\boldsymbol{x}_{0i}\}_{i=1}^{n_0}$, and the pseudo-labels are obtained via the imputation model $\tilde{f}$ trained on the second half of the source dataset $\mathcal{D}_2$. In our vector notation, the target pseudo-labels can be decomposed as in Lemma C.1:

$$
\boldsymbol{X}_0\tilde{\boldsymbol{\theta}} = \boldsymbol{X}_0\boldsymbol{\theta}^* + (\boldsymbol{B}_1 + \boldsymbol{B}_2) + \tilde{\boldsymbol{\varepsilon}}.
$$

We can verify that Assumption B.1 needed to apply the oracle inequality is verified directly per Assumptions 4.1 to 4.4, with $C = RM_2$. The bias $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_2$ is random, which Assumption B.1 allows (Remark B.1): we take $b^2 := 2(\bar{b}_1^2 + \bar{b}_2^2)$, where $\bar{b}_1^2$ and $\bar{b}_2^2$ are the bounds on $\|\boldsymbol{B}_1\|_2^2$ and $\|\boldsymbol{B}_2\|_2^2$ from Lemma C.1. They hold on the same event of probability at least $1 - \delta/2$ as the bound η

on $\left\|\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right\|_\infty$, so item 3 of Assumption B.1 holds without changing the probability budget. Therefore the oracle inequality gives that

$$\sqrt{R_{in}(\hat{f})-R_{in}(f^*)}-\min_{\lambda\in\Lambda}\sqrt{R_{in}(\hat{f}_\lambda)-R_{in}(f^*)}$$

(this intermediate inequality is obtained directly in the proof of the theorem) is smaller up to a constant than

$$\frac{\sqrt{K}}{\sqrt{n_0 k}}\sqrt{K\tilde{\sigma}^2\log^2(\frac{m}{\delta})+Kb^2+\frac{L^2 n_0\eta^4}{k}}, \tag{D.5}$$

with b as above, $\tilde{\sigma}$ is an upper bound on $\|\tilde{\varepsilon}\|_{\psi_1}$, and η is an upper bound on $\left\|\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^*)\right\|_\infty$.

We take the square and consider each term separately, bounding them using Lemma C.1 (for $\boldsymbol{B}_1,\boldsymbol{B}_2,\|\tilde{\varepsilon}\|_{\psi_1}$) and Theorem B.2 (for η).

The variance term:

$$\frac{K}{n_0 k}K\|\tilde{\varepsilon}\|_{\psi_1}^2\log^2(\frac{m}{\delta})\le\frac{K^2\sigma^2}{k^2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\frac{\log^2(m/\delta)}{n_2}.$$

The bias term:

$$\begin{aligned}\frac{K}{n_0 k}Kb^2&=\frac{2K^2}{n_0 k}(\bar{b}_1^2+\bar{b}_2^2)\\&\lesssim\frac{K^2}{k}(\tilde{\lambda}R^2\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|+\frac{L^2M_1^4\sigma^2}{k^2 n_2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\log^7(n_2/\delta))\\&\lesssim\frac{(1+L^2)K^2M_1^4\sigma^2R^2}{k^2}\left\|\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}\right\|\frac{\log^7(n_2/\delta)\log(n_0)}{n_2}.\end{aligned}$$

Finally, the perturbation term:

$$\frac{K}{n_0 k}\frac{L^2 n_0\eta^4}{k}\lesssim\frac{L^2KM_1^6\sigma^2}{k^2}\frac{\log^6(n_2/\delta)}{n_2}.$$

Combining the three terms together yields the desired result.

D.5 Proof of Lemma C.5

We adapt the proof of lemma 6.5 in (Wang, 2026). The only difference is checking the assumptions of their lemma D.5. Use Theorem B.2 that gives for any $\epsilon\in(0,1/2]$

$$\mathbb{P}\left(\left|\left\langle\Phi(\boldsymbol{x}_0),\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\rangle\right|\le cM_1\left(\sigma\log(\frac{n_1}{\epsilon})\frac{\lambda^{-1/4\alpha}}{\sqrt{n_1}}+\lambda^{1/4}\|\boldsymbol{\theta}^*\|\right)\right)\ge1-\epsilon.$$

Plugging in the two bounds on λ, we obtain with probability larger than $1-\epsilon$:

$$\left|\left\langle\Phi(\boldsymbol{x}_0),\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\rangle\right|\lesssim\sqrt{M_1\sigma}\log(\frac{1}{\epsilon})n_1^{-1/4}\log^5(n_1)+g(n_1)^{1/4}\le C\sqrt{M_1\sigma}\log(\frac{1}{\epsilon})$$

for n large enough.

This exponential tail also yields finite moments such that

$$\mathbb{E}\left[\left|\left\langle\Phi(\boldsymbol{x}_0),\hat{\boldsymbol{\theta}}_\lambda-\boldsymbol{\theta}^*\right\rangle\right|^4\right]^{1/4}\le C\sqrt{M_1\sigma}.$$

Thus the inner product satisfies the conditions of lemma 6.5 with $r=C\log(1/\epsilon)\sqrt{M_1\sigma}$ and $\bar{U}=C\sqrt{M_1\sigma}$. The rest follows as in the original proof.

D.6 Proof of Corollary C.1

We recall from Remark A.2 that, since a is k-strongly convex and K-smooth, for every $\lambda \in \Lambda$,

$$\sqrt{\frac{k}{2}} \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\Sigma_0} \leq \sqrt{R(\hat{f}_\lambda) - R(f^*)} \leq \sqrt{\frac{K}{2}} \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\Sigma_0}, \tag{D.6}$$

and

$$\sqrt{\frac{k}{2}} \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\hat{\Sigma}_0} \leq \sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} \leq \sqrt{\frac{K}{2}} \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\hat{\Sigma}_0}. \tag{D.7}$$

For compactness, define

$$c_1 := \sqrt{\frac{k}{2}}, \qquad c_2 := \sqrt{\frac{K}{2}}, \qquad \rho := \frac{c_2}{c_1} = \sqrt{\frac{K}{k}}.$$

Moreover, Lemma C.5 gives, with high probability,

$$\max_{\lambda \in \Lambda} \left| \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\Sigma_0} - \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\hat{\Sigma}_0} \right| \leq x, \tag{D.8}$$

where

$$x = C\sqrt{M_1 \sigma} \sqrt{\frac{\log^2(n_0 m/\delta) \log(m/\delta)}{n_0}}.$$

Finally, Lemma C.4 gives an inequality of the form

$$\sqrt{R_{in}(\hat{f}) - R_{in}(f^*)} \leq \min_{\lambda \in \Lambda} \sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} + b, \tag{D.9}$$

where

$$b = C\sqrt{\hat{\xi}(\tilde{\lambda}, \delta)}.$$

We first compare the in-sample excess risk to the out-of-sample excess risk. For every $\lambda \in \Lambda$, by Eq. (D.7) and Eq. (D.8),

$$\begin{aligned} \sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} &\leq c_2 \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\hat{\Sigma}_0} \\ &\leq c_2 \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\Sigma_0} + c_2 x. \end{aligned}$$

Using the lower bound in Eq. (D.6), we obtain

$$\sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} \leq \rho \sqrt{R(\hat{f}_\lambda) - R(f^*)} + c_2 x. \tag{D.10}$$

Therefore,

$$\min_{\lambda \in \Lambda} \sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} \leq \rho \min_{\lambda \in \Lambda} \sqrt{R(\hat{f}_\lambda) - R(f^*)} + c_2 x. \tag{D.11}$$

We also compare the out-of-sample excess risk of the selected estimator to its in-sample excess risk. By Eq. (D.6) and Eq. (D.8), with $\hat{f} = \hat{f}_{\hat{\lambda}}$,

$$\begin{aligned} \sqrt{R(\hat{f}) - R(f^*)} &\leq c_2 \left\| \hat{\theta}_{\hat{\lambda}} - \boldsymbol{\theta}^* \right\|_{\Sigma_0} \\ &\leq c_2 \left\| \hat{\theta}_{\hat{\lambda}} - \boldsymbol{\theta}^* \right\|_{\hat{\Sigma}_0} + c_2 x. \end{aligned}$$

Using the lower bound in Eq. (D.7), we get

$$\sqrt{R(\hat{f}) - R(f^*)} \leq \rho \sqrt{R_{in}(\hat{f}) - R_{in}(f^*)} + c_2 x. \tag{D.12}$$

Plugging Eq. (D.9) into Eq. (D.12), we obtain

$$\sqrt{R(\hat{f}) - R(f^*)} \leq \rho \left\{ \min_{\lambda \in \Lambda} \sqrt{R_{in}(\hat{f}_\lambda) - R_{in}(f^*)} + b \right\} + c_2 x.$$

Applying Eq. (D.11) yields

$$\begin{aligned}\sqrt{R(\hat{f}) - R(f^*)} &\leq \rho \left\{ \rho \min_{\lambda \in \Lambda} \sqrt{R(\hat{f}_\lambda) - R(f^*)} + c_2 x + b \right\} + c_2 x \\ &= \rho^2 \min_{\lambda \in \Lambda} \sqrt{R(\hat{f}_\lambda) - R(f^*)} + c_2(1+\rho)x + \rho b.\end{aligned}$$

Since $\rho^2 = K/k$, $c_2 = \sqrt{K/2}$, and $\rho = \sqrt{K/k}$, we conclude that

$$\sqrt{R(\hat{f}) - R(f^*)} \leq \frac{K}{k} \min_{\lambda \in \Lambda} \sqrt{R(\hat{f}_\lambda) - R(f^*)} + \sqrt{\frac{K}{2}} \left(1 + \sqrt{\frac{K}{k}} \right) x + \sqrt{\frac{K}{k}}\, b.$$

Substituting the definitions of x and b gives the desired result.

E Proofs of Appendix B.2's pointwise convergence rate

Recall our notations: k is strong convexity constant, M_1 is a uniform bound on the eigenfunctions of the RKHS, M_2 is a uniform bound on the covariates, σ is a bound on the sub-exponential norm of the noise vector, and R is a bound on the true parameter's norm.

Using the isometry between the RKHS and the feature space Eq. (2.4), an equivalent formulation of Eq. (2.7) is

$$\min_{\boldsymbol{\theta} \in \mathbb{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (a(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) - y_i \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|_{\mathbb{H}}^2 \right\}, \tag{E.1}$$

In what follows, denote the regularized empirical GLM loss and its minimizer as:

$$L_\lambda(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (a(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) - y_i \langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) + \frac{\lambda}{2} \|\boldsymbol{\theta}\|^2 \,, \quad \hat{\boldsymbol{\theta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \mathbb{H}} L_\lambda(\boldsymbol{\theta}),$$

with gradient

$$\nabla L_\lambda(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (a'(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) - y_i) \Phi(\boldsymbol{x}_i) + \lambda \boldsymbol{\theta} \,, \quad \nabla L_\lambda(\boldsymbol{\theta}^*) = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda \boldsymbol{\theta}^*,$$

and Hessian

$$\nabla^2 L_\lambda(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} a''(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) (\Phi(\boldsymbol{x}_i) \otimes \Phi(\boldsymbol{x}_i)) + \lambda \boldsymbol{I}.$$

We need the following definition in the rest of the proof:

Definition E.1 (Strong convexity region)**.** Define $\Omega(\rho) := \{\boldsymbol{\theta} : \|\boldsymbol{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|_\infty \leq \rho C\}$ the domain over which the local strong convexity implies that $\forall \boldsymbol{\theta} \in \Omega(\rho)$, $\forall i \in [n], a''(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} \rangle) \geq k_\rho$. Denote also $k := k_1$, the smallest strong convexity constant when $\rho \leq 1$.

Remark E.1. It follows that $\forall \boldsymbol{\theta} \in \Omega(\rho), \nabla^2 L_\lambda(\boldsymbol{\theta}) \succeq k_\rho \hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}$.

To stay more general in this section, we will assume that the kernel eigenvalues in the $L^2(\mathcal{P}, \mathcal{X})$ basis have a polynomial decay of exponent $\alpha \geq 1$: $\forall j, \mu_j \leq A j^{-2\alpha}$. Only at the end will we replace α by 1, as per our Assumption 4.1.

GLM estimator properties To characterize the estimation error of the estimator, we let $\boldsymbol{\theta}(t) = \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)$ and write the Taylor expansion of ∇L_λ between $\hat{\boldsymbol{\theta}}_\lambda$ and $\boldsymbol{\theta}^*$:

$$0 = \nabla L_\lambda(\hat{\boldsymbol{\theta}}_\lambda) = \nabla L_\lambda(\boldsymbol{\theta}^*) + \left[\int_0^1 \nabla^2 L_\lambda(\boldsymbol{\theta}(t)) dt \right] \cdot (\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*).$$

Then take the inverse to get an expression of:

$$\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = -\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X} + \lambda \boldsymbol{I}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda \boldsymbol{\theta}^*\right), \tag{E.2}$$

where the diagonal matrix $\boldsymbol{D}$ has components

$$D_{ii} = \int_0^1 a''\left(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}(t)\rangle\right) dt.$$

However, to turn this non-linear equation into a usable expression, we need to know that $\boldsymbol{\theta}(t)$ is in the region of strong convexity to lower bound $\boldsymbol{D}$ by $k\boldsymbol{I}$. This motivates our first lemma which proves a crude consistency rate with high probability. This coarse-grained analysis will serve as a stepping stone to prove the main theorem.

E.1 Coarse-grained analysis of the estimator consistency

Lemma E.1. *Let $g(n) = \frac{1}{\log^\beta(n)} \underset{n\to\infty}{\longrightarrow} 0, \quad \beta > 0$, and choose $\delta \in (0, 1/7]$. Then, under Assumptions 2.1 and 4.1 to 4.4, there exists constants $c, c_0, c_1 > 0$ and $c_2 \geq k$, such that when $\rho = c_0 g(n)^{1/4}$ and:*

$$\begin{cases} \lambda \leq \frac{c_1 g(n)}{R^4} = \frac{c_1}{R^4 \log^\beta(n)} \\ \lambda \geq c_2 \left(\frac{M_1^2 \sigma \sqrt{\log(4/\delta)}\log n}{\sqrt{n}}\right)^2 \frac{1}{g(n)^{1/2}} = \frac{c_2 M_1^2 \sigma^2 \log(\frac{4}{\delta})(\log n)^{2+\beta/2}}{n} \end{cases}$$

We have:

$$\mathbb{P}\left(\hat{\boldsymbol{\theta}}_\lambda \in \mathring{\Omega}(\rho)\right) \geq 1 - \delta,$$

namely with probability at least $1-\delta$,

$$\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty \leq c g(n)^{1/4}$$

Proof. By contradiction, assume that $\hat{\boldsymbol{\theta}}_\lambda \notin \Omega(\rho)$. We need show that $\forall \boldsymbol{\theta} \notin \Omega(\rho), L_\lambda(\boldsymbol{\theta}) > L_\lambda(\boldsymbol{\theta}^*)$, for a well-chosen ρ, thus contradicting the optimality of $\hat{\boldsymbol{\theta}}_\lambda$. Inside $\Omega(\rho)$, we can use the local k_ρ-strong convexity. In this proof, we denote $\boldsymbol{S}_{\lambda,\rho} := k_\rho \hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}$.

Let the ellipsoid $E_\rho = \left\{\boldsymbol{\theta} : \|\boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq 2\,\|\boldsymbol{S}_{\lambda,\rho}^{-1/2}\nabla L_\lambda(\boldsymbol{\theta}^*)\|\right\}$. We begin by proving the two following facts:

- **Fact 1:** $\forall \boldsymbol{\theta} \in \mathring{\Omega}(\rho) \setminus E_\rho, L_\lambda(\boldsymbol{\theta}) > L_\lambda(\boldsymbol{\theta}^*)$.

  Proof. Let $\boldsymbol{\theta} \in \mathring{\Omega}(\rho)$. Then we can apply Taylor Lagrange between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ with a $\boldsymbol{\theta}_{in} = \boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \in \mathring{\Omega}(\rho)$ (for some $t \in (0,1)$) on the line segment connecting the two vectors. We can then use the strong convexity of a to write:

  $$\begin{aligned} L_\lambda(\boldsymbol{\theta}) &= L_\lambda(\boldsymbol{\theta}^*) + \langle \nabla L_\lambda(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \frac{1}{2}\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \nabla^2 L_\lambda(\boldsymbol{\theta}_{in})(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right)\rangle \\ &\geq L_\lambda(\boldsymbol{\theta}^*) + \langle \nabla L_\lambda(\boldsymbol{\theta}^*), \boldsymbol{\theta} - \boldsymbol{\theta}^*\rangle + \frac{1}{2}\left\langle \boldsymbol{\theta} - \boldsymbol{\theta}^*, \boldsymbol{S}_{\lambda,\rho}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right)\rangle, \end{aligned}$$

  i.e.

  $$L_\lambda(\boldsymbol{\theta}) \geq L_\lambda(\boldsymbol{\theta}^*) + \left\langle \boldsymbol{S}_{\lambda,\rho}^{-1/2}\nabla L_\lambda(\boldsymbol{\theta}^*), \boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\rangle + \frac{1}{2}\left\|\boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\|^2.$$

  Using Cauchy-Schwarz, we get:

  $$L_\lambda(\boldsymbol{\theta}) - L_\lambda(\boldsymbol{\theta}^*) \geq \left\|\boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\|\left(\frac{1}{2}\left\|\boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right\| - \left\|\boldsymbol{S}_{\lambda,\rho}^{-1/2}\nabla L_\lambda(\boldsymbol{\theta}^*)\right\|\right),$$

  which yields the desired result. □

- **Fact 2:** If
$$\rho > \frac{2}{C} \max_i \left\{ \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\| \right\}, \tag{E.3}$$
then E_ρ is a proper subset of $\Omega(\rho)$.

Proof. Let $\boldsymbol{\theta} \in E_\rho, i \in [n]$. Then:
$$|\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle| = \left| \left\langle \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i), \boldsymbol{S}_{\lambda,\rho}^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\rangle \right| \leq \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| \boldsymbol{S}_{\lambda,\rho}^{1/2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\|.$$
By definition of E_ρ, we get:
$$|\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta} - \boldsymbol{\theta}^* \rangle| \leq 2 \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\|.$$
Thus, if $\rho > \frac{2}{C} \max_i \left\{ \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\| \right\}$, $\boldsymbol{\theta} \in \mathring{\Omega}(\rho)$, and the result is proved. □

Together these two facts imply either that $\hat{\boldsymbol{\theta}}_\lambda \notin \Omega(\rho)$ or that $\hat{\boldsymbol{\theta}}_\lambda \in E_\rho \subset \mathring{\Omega}(\rho)$ by optimality of $\hat{\boldsymbol{\theta}}_\lambda$. Let us prove that the first possibility is actually impossible.

Let $\boldsymbol{\theta} \notin \Omega(\rho)$ and construct the line parametrized by $\boldsymbol{\theta}(t) = \boldsymbol{\theta}^* + t(\boldsymbol{\theta} - \boldsymbol{\theta}^*), t \in [0,1]$ linking $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$. By continuity, it intersects $\Omega(\rho)$ and we define the intersection point as $t_b \in (0,1), \boldsymbol{\theta}(t_b) \in \partial\Omega(\rho)$. Given that $\boldsymbol{\theta}(t_b) \notin E_\rho$, we know from **Fact 1** that $L_\lambda(\boldsymbol{\theta}(t_b)) \geq L_\lambda(\boldsymbol{\theta}^*)$. Then by convexity of L_λ (true everywhere), $L_\lambda(\boldsymbol{\theta}) \geq \frac{1}{t_b} L_\lambda(\boldsymbol{\theta}(t_b)) + (1 - \frac{1}{t_b}) L_\lambda(\boldsymbol{\theta}^*) > L_\lambda(\boldsymbol{\theta}^*)$.

Finally, we have proved that with a well chosen ρ satisfying Eq. (E.3), $\hat{\boldsymbol{\theta}}_\lambda \in E_\rho \subset \mathring{\Omega}(\rho)$.

Let us now prove that the inequality (E.3) is feasible and extract the adequate ρ. This is a nonlinear inequality in ρ. However, we know that k_ρ is decreasing as ρ increases (since ρ expands the region of strong convexity), which makes the RHS of (E.3) increasing in ρ. If we can assume that ρ is upper bounded, say $\rho \leq 1$, and then can show the existence of $\rho > \frac{2}{C} \max_i \left\{ \left\| S_{\lambda,1}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| S_{\lambda,1}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\| \right\}$ still smaller than 1, we are done. We are going to prove that it is possible by computing explicitly the RHS for $\rho = 1$:

Computation of the RHS of (E.3) **for** $\rho = 1$
$$\begin{aligned} RHS &= \frac{2}{C} \max_i \left\{ \left\| S_{\lambda,1}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| S_{\lambda,1}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\| \right\} \\ &= \frac{2}{C} \max_i \left\{ \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}))^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} (\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda \boldsymbol{\theta}^*)) \right\| \right\} \\ &\leq \frac{2}{C} \max_i \left\{ \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}))^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left(\frac{1}{n} \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\| + \lambda \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} \boldsymbol{\theta}^* \right\| \right) \right\}, \end{aligned} \tag{E.4}$$
where $k = k_1$.

Define the event
$$\mathcal{A} := \left\{ \frac{1}{2}(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \preceq k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I} \preceq \frac{3}{2}(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \right\}.$$
By Lemma G.3 applied with $\gamma = 1/2$, $\lambda = \lambda/k$ and $\delta = \delta/2$, which relates the sample and population covariance, since $\delta \in (0, 1/7]$ and our choice of λ satisfies $\lambda \geq \frac{CM_2^2 \log(2n/\delta)}{n}$ where $C \geq k > 0$ for n large enough, we have
$$\mathbb{P}(\mathcal{A}) \geq 1 - \frac{\delta}{2}.$$
We consider the three norm terms of Eq. (E.4) successively. We study the square norm to use inner products and then take the square-root. $\lesssim$ only hides universal multiplicative constants.

- Under $\mathcal{A}$, $\left\langle \Phi(\boldsymbol{x}_i), (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \Phi(\boldsymbol{x}_i) \right\rangle \leq 2 \left\langle \Phi(\boldsymbol{x}_i), (k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \Phi(\boldsymbol{x}_i) \right\rangle$. We make use of Lemma G.1 to get: $\left\langle \Phi(\boldsymbol{x}_i), (k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \Phi(\boldsymbol{x}_i) \right\rangle \lesssim M_1^2 \lambda^{-1/2\alpha}$. Therefore, taking the square-root, we obtain

$$\left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \lesssim M_1 \lambda^{-1/4\alpha}.$$

- Under $\mathcal{A}$, the bias term writes: $\lambda \left\| (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} \boldsymbol{\theta}^* \right\| \lesssim \lambda \left\| (k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2} \boldsymbol{\theta}^* \right\|$. Then, using $(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \preceq \boldsymbol{I}/\lambda$,

$$\left\| (k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2} \lambda \boldsymbol{\theta}^* \right\| \leq \sqrt{\lambda} \left\| \boldsymbol{\theta}^* \right\| \lesssim \lambda^{1/2}$$

  for bounded $\boldsymbol{\theta}^*$.

- For the squared variance term $\left\langle \varepsilon, \frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\rangle$, conditioned on $\boldsymbol{X}$, we use the quadratic form tail bound proved in Lemma G.2 for the sub-exponential random vector ε and PSD matrix $A = \frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top$, which bounds with high probability by the trace of the PSD operator. Since the δ defined earlier verifies $\delta \leq 1/7 \leq 4/e$, we have:

$$\mathbb{P}\left(\left\langle \varepsilon, \frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\rangle \lesssim \sigma^2 \log^2(n) \log(4/\delta) \operatorname{Tr}\left(\frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \right) \,\middle|\, \boldsymbol{X} \right) \geq 1 - \frac{\delta}{2},$$

  and thus integrating over the whole event:

$$\mathbb{P}\left(\left\langle \varepsilon, \frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\rangle \lesssim \sigma^2 \log^2(n) \log(4/\delta) \operatorname{Tr}\left(\frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \right) \,\middle|\, \mathcal{A} \right) \geq 1 - \frac{\delta}{2}.$$

  Now, under $\mathcal{A}$ according to Lemma G.3,

$$\operatorname{Tr}\left(\frac{1}{n^2} \boldsymbol{X} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^\top \right) = \frac{1}{n} \operatorname{Tr}\left((k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}} \right) \leq \frac{2M_1^2}{n} \operatorname{Tr}\left((k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma} \right).$$

  Then applying again Lemma G.1, we have that

$$\operatorname{Tr}\left((k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma} \right) = r(\lambda/k)/k \lesssim \lambda^{-1/2\alpha}.$$

  Therefore:

$$\mathbb{P}\left(\left\| \frac{1}{n} (k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1/2} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\| \lesssim \sigma M_1 \sqrt{\log(4/\delta)} n^{-1/2} \lambda^{-1/4\alpha} \log n \,\middle|\, \mathcal{A} \right) \geq 1 - \frac{\delta}{2}.$$

All in all, since our analysis holds for any i, there exists $c > 0$, such that the event

$$\mathcal{E} := \left\{ RHS \leq c M_1 \lambda^{-1/4\alpha} \left(\lambda^{1/2} \left\| \boldsymbol{\theta}^* \right\| + \sigma M_1 \sqrt{\log(4/\delta)} \lambda^{-1/4\alpha} \frac{\log n}{\sqrt{n}} \right) \right\}$$

has conditional probability

$$\mathbb{P}\left(\mathcal{E} \,\middle|\, \mathcal{A} \right) \geq 1 - \frac{\delta}{2},$$

and

$$\mathbb{P}\left(\mathcal{E} \right) \geq \mathbb{P}\left(\mathcal{E} \,\middle|\, \mathcal{A} \right) \mathbb{P}\left(\mathcal{A} \right) \geq \left(1 - \frac{\delta}{2} \right)^2 \geq 1 - \delta.$$

Crude consistency rate. Therefore, if we set

$$\left(\sigma^2 M_1^2 \log(\frac{4}{\delta})\right)^{\alpha} \left(\frac{\log n}{\sqrt{n}}\right)^{2\alpha} \frac{1}{g(n)^{\alpha/2}} \lesssim \lambda \lesssim \frac{g(n)}{R^4} \tag{E.5}$$

for a slowly decaying $g(n) \to 0$ (and still respecting the condition of event $\mathcal{A}$ [4]), since $1/2 - 1/4\alpha \geq 1/4$, we have with probability at least $1-\delta$:

$$RHS \lesssim g(n)^{1/4} \to 0.$$

We can simplify inequality (E.5) since it is still verified when setting $\alpha = 1$, and with $g(n) = \frac{1}{\log^{\beta}(n)}, \beta > 0$, so that it becomes

$$\frac{M_1^2 \sigma^2 \log(\frac{4}{\delta})(\log n)^{2+\beta/2}}{n} \lesssim \lambda \lesssim \frac{1}{R^4 \log^{\beta}(n)}. \tag{E.6}$$

Namely, if we set $\rho = g(n)^{1/4}$, inequality (E.3) is verified and $\hat{\boldsymbol{\theta}}_\lambda \in \Omega(\rho)$ so that the following consistency rate is established with probability at least $1-\delta$:

$$\left\| \boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\|_\infty \lesssim g(n)^{1/4} = \frac{1}{\log^{\beta/4}(n)}.$$

□

E.2 Proof of Theorem B.2

Building on the previous lemma, we refine the analysis of the consistency rate in the theorem.

Theorem B.2 (Pointwise convergence rate)**.** *Consider a ridge-regularized kernel GLM estimator $\hat{f}_\lambda$ trained on n i.i.d. samples from the source distribution $\mathcal{P}$ via Eq.* (2.7)*. Let $\beta \in (0, 8)$, and choose $\delta \in (0, 1/e]$. Then, under Assumptions 2.1 and 4.1 to 4.4, there exist constants $c, c', c_1, c_2 > 0$, such that for any given $\boldsymbol{x}_0 \in \mathrm{supp}(\mathcal{P})$, when:*

$$\begin{cases} \lambda \leq \frac{c_1}{R^4 \log^{\beta}(n)} \\ \lambda \geq \frac{c_2 M_1^2 \sigma^2 \log(\frac{8}{\delta})(\log n)^{2+\beta/2}}{n} \end{cases},$$

the following inequality holds with probability at least $1-\delta$:

$$\left| \hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0) \right| \leq c M_1 \left(\sigma \log(\frac{n}{\delta}) \frac{\lambda^{-1/4}}{\sqrt{n}} + \lambda^{1/4} \left\| f^* \right\|_{\mathcal{F}} \right). \tag{B.11}$$

When setting λ to its lower bound, the result becomes

$$\left| \hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0) \right| \leq c' \sqrt{M_1^3 \sigma} n^{-1/4} \left(\log(\frac{n}{\delta}) \right)^3. \tag{B.12}$$

Remark E.2. When union bounding on all the $\boldsymbol{x}_0$ of a target $\boldsymbol{X}_0$ of length n_0, we get with $g(n) := \frac{1}{\log^{\beta}(n)}$ that when

$$\lambda \asymp \left(\frac{M_1 \sigma \sqrt{\log(8n_0/\delta)} \log n}{\sqrt{n}} \right)^2 \frac{1}{g(n)^{1/2}},$$

$$\mathbb{P}\left(\left\| \boldsymbol{X}_0(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\|_\infty \lesssim \sqrt{M_1^3 \sigma} n^{-1/4} (\log(\frac{n}{\delta}))^3 \right) \geq 1 - \delta.$$

Proof. Recall Eq. (E.2):

$$\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = -\left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{D} \boldsymbol{X} + \lambda \boldsymbol{I} \right)^{-1} \left(\frac{1}{n} \boldsymbol{X}^\top \boldsymbol{\varepsilon} + \lambda \boldsymbol{\theta}^* \right)$$

[4] For instance, we take $1/g(n)$ to be a polylog

where the diagonal matrix $\boldsymbol{D}$ has components

$$D_{ii} = \int_0^1 a''\left(\left\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\rangle\right) dt.$$

We denote $\boldsymbol{H} := \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{D}\boldsymbol{X} + \lambda \boldsymbol{I}$ in what follows.

Since $\boldsymbol{D}$ and $\boldsymbol{H}$ depend on $\hat{\boldsymbol{\theta}}_\lambda$, they are random and correlated with ε, so let us consider their deterministic (in fixed-design) versions $\bar{D}_{ii} := a''\left(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^*\rangle\right)$ and $\bar{\boldsymbol{H}} := \frac{1}{n}\boldsymbol{X}^\top \bar{\boldsymbol{D}}\boldsymbol{X} + \lambda \boldsymbol{I}$. Let $\boldsymbol{x}_0 \in \operatorname{supp}(P)$ and $\Phi_0 := \Phi(\boldsymbol{x}_0)$ the associated vector in the RKHS. We can now write a sort of bias-variance decomposition plus perturbation from (E.2):

$$\left|\left\langle \Phi_0, \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\rangle\right| \le \lambda \left|\langle \Phi_0, \boldsymbol{H}^{-1}\boldsymbol{\theta}^*\rangle\right| + \left|\left\langle \Phi_0, \frac{1}{n}\bar{\boldsymbol{H}}^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\right\rangle\right| + \left|\left\langle \Phi_0, \frac{1}{n}(\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1})\boldsymbol{X}^\top \boldsymbol{\varepsilon}\right\rangle\right|. \tag{E.7}$$

We will show that the perturbation term is small.

We let $k = k_1$ be the constant of strong convexity of the link function a. This constant is the smallest allowed for the range of ρ considered, i.e. associated to the largest region of strong convexity. a is thus k-strongly convex on $\Omega(\rho), \forall \rho \le 1$. As such, $\bar{\boldsymbol{D}} \succeq k\boldsymbol{I}$ and $\bar{\boldsymbol{H}} \succeq k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}$, and if we let the event of Lemma E.1 be:

$$\mathcal{B} := \left\{ \left\| \boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty \lesssim g(n)^{1/4} \right\},$$

which holds with probability larger than $1 - \delta/2$, when

$$\begin{cases} \lambda \le \frac{c_1 g(n)}{\|\boldsymbol{\theta}^*\|^4} \\ \lambda \ge c_2 \left(\frac{M_1 \sigma \sqrt{\log(8/\delta)} \log n}{\sqrt{n}} \right)^{2\alpha} \frac{1}{g(n)^{\alpha/2}}, \end{cases}$$

and $\delta \in (0, 2/7]$, we also have a fortiori under $\mathcal{B}$ that $\boldsymbol{D} \succeq k\boldsymbol{I}$ and $\boldsymbol{H} \succeq k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}$. Recall as well that under $\mathcal{B}$, the events from Lemma E.1

$$\mathcal{A}_1 = \left\{ \frac{1}{2}(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \preceq k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I} \preceq \frac{3}{2}(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \right\},$$

and

$$\mathcal{A}_2 = \left\{ \left\| \frac{1}{\sqrt{n}} \left(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}\right)^{-1/2} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\| \lesssim M_1 \sigma \sqrt{\log(8/\delta)} \log(n) \lambda^{-1/4\alpha} \right\}$$

hold as well. **We place ourselves under that event $\mathcal{B}$ in the rest of the proof**, everything is conditional on $\boldsymbol{X}$ in particular. We study the three terms of (E.7) separately. Begin by noting the following fact.

Fact E.1 (Relative error)**.**

$$\forall i \in [n], \frac{\left|D_{ii} - \bar{D}_{ii}\right|}{\bar{D}_{ii}} \le \frac{1}{2}$$

for n large enough, the rank depending on the Lipschitz and strong convexity constant of the log-partition function a [5]*. It implies successively the following inequalities:*

$$-\frac{1}{2}\bar{\boldsymbol{H}} \preceq \boldsymbol{H} - \bar{\boldsymbol{H}} \preceq \frac{1}{2}\bar{\boldsymbol{H}}, \quad \|\boldsymbol{\Delta}\| \le \frac{1}{2},$$

where $\boldsymbol{\Delta} := \bar{\boldsymbol{H}}^{-1/2}\boldsymbol{H}\bar{\boldsymbol{H}}^{-1/2} - I$. We also get that $\|\boldsymbol{\Delta}\| \le \frac{L}{2k}\left\| \boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty$.

[5] E.g. for $g(n) = 1/\log^4(n)$, with the logistic regression loss, n must be larger than 28. Compute other examples with explicit constants!

Proof.

$$\begin{aligned}
\left|D_{ii} - \bar{D}_{ii}\right| &= \left| \int_0^1 \left(a'' \left(\left\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\rangle \right) - a'' \left(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^* \rangle \right) \right) dt \right| \\
&\leq \int_0^1 \left| a'' \left(\left\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\rangle \right) - a'' \left(\langle \Phi(\boldsymbol{x}_i), \boldsymbol{\theta}^* \rangle \right) \right| dt \\
&\leq \int_0^1 Lt \left| \left\langle \Phi(\boldsymbol{x}_i), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle \right| dt \quad \text{by Lipschitz continuity of } a'' \\
&= \frac{L}{2} \left| \left\langle \Phi(\boldsymbol{x}_i), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle \right| \\
&\lesssim Lg(n)^{1/4} \quad \text{under event } \mathcal{B}.
\end{aligned}$$

Since we also have $\bar{D}_{ii} \geq k$, we deduce

$$\frac{\left|D_{ii} - \bar{D}_{ii}\right|}{\bar{D}_{ii}} \lesssim \frac{L}{k} g(n)^{1/4} \leq \frac{1}{2}$$

for n large enough.

Since they are diagonal matrices, it implies that

$$-\frac{1}{2}\bar{\boldsymbol{D}} \preceq \boldsymbol{D} - \bar{\boldsymbol{D}} \preceq \frac{1}{2}\bar{\boldsymbol{D}},$$

and thus

$$-\frac{1}{2}\bar{\boldsymbol{H}} \preceq \boldsymbol{H} - \bar{\boldsymbol{H}} \preceq \frac{1}{2}\bar{\boldsymbol{H}},$$

and

$$-\frac{1}{2}\boldsymbol{I} \preceq \boldsymbol{\Delta} \preceq \frac{1}{2}\boldsymbol{I},$$

hence $\|\boldsymbol{\Delta}\| \leq \frac{1}{2}$.

Also note that if we denote by $\boldsymbol{E} = \bar{\boldsymbol{D}}^{-1/2}(\boldsymbol{D} - \bar{\boldsymbol{D}})\bar{\boldsymbol{D}}^{-1/2}$ the diagonal matrix of elements $\frac{D_{ii} - \bar{D}_{ii}}{\bar{D}_{ii}}$, and define the matrix $\boldsymbol{A} = \frac{1}{\sqrt{n}}\bar{\boldsymbol{D}}^{1/2}\boldsymbol{X}\bar{\boldsymbol{H}}^{-1/2}$, then

$$\boldsymbol{\Delta} = \boldsymbol{A}^\top \boldsymbol{E} \boldsymbol{A}\,, \quad \text{and } \boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I} - \lambda \bar{\boldsymbol{H}} \preceq \boldsymbol{I},$$

so that

$$\|\boldsymbol{\Delta}\| \leq \|\boldsymbol{E}\| \leq \frac{L}{2k} \left\| \boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*) \right\|_\infty \lesssim g(n)^{1/4}.$$

□

Note also that the strong convexity property allows to lower bound $\boldsymbol{H}, \bar{\boldsymbol{H}}$ by $k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I}$, so that using successively lemmas G.3, G.1, and Lemma G.2 for the third point, we get the following fact:

Fact E.2. *The matrix $\boldsymbol{A}$ can be either $\boldsymbol{H}$ or $\bar{\boldsymbol{H}}$*

- $\left\| \boldsymbol{A}^{-1/2} \Phi_0 \right\| \lesssim M_1 \lambda^{-1/4\alpha}$.
- $\left\| \frac{1}{\sqrt{n}} \boldsymbol{A}^{-1/2} \boldsymbol{X}^\top \right\| \lesssim 1$.
- $\left\| \frac{1}{\sqrt{n}} \boldsymbol{A}^{-1/2} \boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\| \lesssim M_1 \sigma \sqrt{\log(8/\delta)} \lambda^{-1/4\alpha} \log n$. [6]

[6] Thanks to the symmetric form, we can bound $\boldsymbol{H}$ and $\bar{\boldsymbol{H}}$ immediately, which makes the use of the subexponential tail bound valid, no need to wait for taking the trace

Proof. For the first item, write:

$$\begin{aligned}\left\|\boldsymbol{A}^{-1/2}\Phi_0\right\|^2 &= \left\langle \Phi_0, \boldsymbol{A}^{-1}\Phi_0\right\rangle \\ &\le \left\langle \Phi_0, (k\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1}\Phi_0\right\rangle \quad \text{under } \mathcal{B} \\ &\le 2\left\langle \Phi_0, (k\boldsymbol{\Sigma}+\lambda\boldsymbol{I})^{-1}\Phi_0\right\rangle \quad \text{via Lemma G.3} \\ &\lesssim M_1^2\lambda^{-1/2\alpha} \quad \text{via Lemma G.1.}\end{aligned}$$

For the second item, we have:

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\boldsymbol{A}^{-1/2}\boldsymbol{X}^\top\right\|^2 &= \frac{1}{n}\left\|\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^\top\right\| \\ &\le \frac{1}{n}\left\|\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^\top\right\| \quad \text{under } \mathcal{B} \\ &= \left\|(k\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}\right\| \\ &\lesssim 1.\end{aligned}$$

And for the third item:

$$\begin{aligned}\left\|\frac{1}{\sqrt{n}}\boldsymbol{A}^{-1/2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\| &\le \left\|\frac{1}{\sqrt{n}}(k\hat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I})^{-1/2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\| \quad \text{under } \mathcal{B} \\ &\le M_1\sigma\sqrt{\log(8/\delta)}\lambda^{-1/4\alpha}\log n \quad \text{directly from event } \mathcal{A}_2.\end{aligned}$$

□

We also have the simple bound where $\boldsymbol{A}$ can be either $\boldsymbol{H}$ or $\bar{\boldsymbol{H}}$:

$$\boldsymbol{A}^{-1/2} \preceq \lambda^{-1/2}\boldsymbol{I}.$$

Computation of the bias term of (E.7)

$$\begin{aligned}\lambda\left|\left\langle \Phi_0, \boldsymbol{H}^{-1}\boldsymbol{\theta}^*\right\rangle\right| &= \lambda\left|\left\langle \boldsymbol{H}^{-1/2}\Phi_0, \boldsymbol{H}^{-1/2}\boldsymbol{\theta}^*\right\rangle\right| \\ &\le \lambda\left\|\boldsymbol{H}^{-1/2}\Phi_0\right\|\cdot\left\|\boldsymbol{H}^{-1/2}\boldsymbol{\theta}^*\right\| \quad \text{(by Cauchy-Schwarz)} \\ &\lesssim \lambda\cdot M_1\lambda^{-1/4\alpha}\cdot\lambda^{-1/2}\left\|\boldsymbol{\theta}^*\right\| \quad \text{(by Fact E.2)} \\ &= M_1\lambda^{(2-1/\alpha)/4}\left\|\boldsymbol{\theta}^*\right\| \\ &\le M_1\lambda^{1/4}\left\|\boldsymbol{\theta}^*\right\| \quad \text{since } \alpha\ge 1 \text{ and } \lambda \text{ is small.}\end{aligned}$$

Computation of the variance term of (E.7) Conditionally on $\boldsymbol{X}$, $\boldsymbol{\varepsilon}$ is sub-exponential vector with $\|\boldsymbol{\varepsilon}\|_{\psi_1}\le\sigma$. Hence, automatically for any vector $\boldsymbol{a}\in\mathbb{R}^n$, $\langle\boldsymbol{a},\boldsymbol{\varepsilon}\rangle$ is a sub-exponential variable with $\|\langle\boldsymbol{a},\boldsymbol{\varepsilon}\rangle\|_{\psi_1}\le\|a\|_2\,\sigma$, such that the following tail bound holds $\forall\delta'\in(0,1)$ with probability at least $1-\delta'$:

$$|\langle\boldsymbol{a},\boldsymbol{\varepsilon}\rangle| \le \sigma\,\|a\|_2\log(2/\delta').$$

Therefore, the variance term can be bound conditionally on the event $\mathcal{B}$ with probability at least $1-\delta'$, as:

$$\begin{aligned}\left|\left\langle \Phi_0, \frac{1}{n}\bar{\boldsymbol{H}}^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\rangle\right| &\le \sigma\log(2/\delta')\frac{1}{n}\left\|\boldsymbol{X}\bar{\boldsymbol{H}}^{-1}\Phi_0\right\|_2 \\ &\le \sigma\log(2/\delta')\frac{1}{\sqrt{n}}\left\|\bar{\boldsymbol{H}}^{-1/2}\Phi_0\right\|\cdot\left\|\frac{1}{\sqrt{n}}\bar{\boldsymbol{H}}^{-1/2}\boldsymbol{X}^\top\right\| \\ &\lesssim \sigma\log(2/\delta')\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha}\cdot 1 \quad \text{(by Fact E.2)} \\ &= \sigma\log(2/\delta')\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha}.\end{aligned}$$

Computation of the perturbation term of (E.7) First note that $(\boldsymbol{I}+\boldsymbol{\Delta})^{-1} = \boldsymbol{I} - \boldsymbol{\Delta}(\boldsymbol{I}+\boldsymbol{\Delta})^{-1}$ and

$$\begin{aligned}\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1} &= \bar{\boldsymbol{H}}^{-1/2}\left[\left(\bar{\boldsymbol{H}}^{-1/2}\boldsymbol{H}\bar{\boldsymbol{H}}^{-1/2}\right)^{-1} - \boldsymbol{I}\right]\bar{\boldsymbol{H}}^{-1/2} \\ &= \bar{\boldsymbol{H}}^{-1/2}\left[(\boldsymbol{I}+\boldsymbol{\Delta})^{-1} - \boldsymbol{I}\right]\bar{\boldsymbol{H}}^{-1/2} \\ &= -\bar{\boldsymbol{H}}^{-1/2}\left[\boldsymbol{\Delta}\,(\boldsymbol{I}+\boldsymbol{\Delta})^{-1}\right]\bar{\boldsymbol{H}}^{-1/2},\end{aligned}$$

and since $\|\boldsymbol{\Delta}\| \leq 1/2$,

$$\left\|(\boldsymbol{I}+\boldsymbol{\Delta})^{-1}\right\| \leq \sum_{n=0}^{\infty}\|\boldsymbol{\Delta}\|^n = \frac{1}{1-\|\boldsymbol{\Delta}\|} \leq 2$$

hence the perturbation term becomes:

$$\begin{aligned}\left|\left\langle \Phi_0, \frac{1}{n}(\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1})\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\rangle\right| &= \left|\left\langle \Phi_0, \frac{1}{n}\left[\bar{\boldsymbol{H}}^{-1/2}\left[\boldsymbol{\Delta}\,(\boldsymbol{I}+\boldsymbol{\Delta})^{-1}\right]\bar{\boldsymbol{H}}^{-1/2}\right]\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\rangle\right| \\ &\leq \left\|\bar{\boldsymbol{H}}^{-1/2}\Phi_0\right\| \cdot 2\,\|\boldsymbol{\Delta}\| \cdot \frac{1}{\sqrt{n}}\left\|\frac{1}{\sqrt{n}}\bar{\boldsymbol{H}}^{-1/2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right\| \\ &\lesssim M_1\lambda^{-1/4\alpha}\cdot\|\boldsymbol{\Delta}\|\cdot\frac{1}{\sqrt{n}}M_1\sigma\sqrt{\log(8/\delta)}\lambda^{-1/4\alpha}\log n \\ &= \frac{\log n}{\sqrt{n\lambda^{1/\alpha}}}\,\|\boldsymbol{\Delta}\|\,M_1^2\sigma\sqrt{\log(8/\delta)}.\end{aligned}$$

So far, we have proved that there exists a constant $c>0$ such that for any $\boldsymbol{x}_0 \in \mathrm{supp}(P)$:

$$\mathbb{P}\left(\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\rangle\right| \leq c\left(\frac{\log n}{\sqrt{n\lambda^{1/\alpha}}}\,\|\boldsymbol{\Delta}\|\,M_1^2\sigma\sqrt{\log(\frac{8}{\delta})} + \sigma\log(\frac{2}{\delta'})\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha} + M_1\lambda^{1/4}\,\|\boldsymbol{\theta}^*\|\right)\,\middle|\,\mathcal{B}\right) \geq 1-\delta'.$$

In particular, by union bound on all the $\boldsymbol{x}_i$ and $\boldsymbol{x}_0$, we have simultaneously with probability larger than $1-(n+1)\delta'$ conditionally on $\mathcal{B}$:

$$\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty \leq c\left(\frac{\log n}{\sqrt{n\lambda^{1/\alpha}}}\,\|\boldsymbol{\Delta}\|\,M_1^2\sigma\sqrt{\log(\frac{8}{\delta})} + \sigma\log(2/\delta')\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha} + M_1\lambda^{1/4}\,\|\boldsymbol{\theta}^*\|\right), \tag{E.8}$$

and

$$\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*\right\rangle\right| \leq c\left(\frac{\log n}{\sqrt{n\lambda^{1/\alpha}}}\,\|\boldsymbol{\Delta}\|\,M_1^2\sigma\sqrt{\log(\frac{8}{\delta})} + \sigma\log(\frac{2}{\delta'})\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha} + M_1\lambda^{1/4}\,\|\boldsymbol{\theta}^*\|\right). \tag{E.9}$$

But by construction from Fact E.1, we also know that $\|\boldsymbol{\Delta}\| \leq C_1\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty$ where $C_1>0$ is a constant depending on the Lipschitz and strong convexity constant.

We can plug this expression in Eq. (E.8) and regroup the terms to get:

$$\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty\left(1 - \frac{cC_1\log n}{\sqrt{n\lambda^{1/\alpha}}}M_1^2\sigma\sqrt{\log(\frac{8}{\delta})}\right) \leq cM_1\left(\sigma\log(2/\delta')\frac{1}{\sqrt{n}}\lambda^{-1/4\alpha} + \lambda^{1/4}\,\|\boldsymbol{\theta}^*\|\right).$$

Thus, if $\frac{cC_1\log n}{\sqrt{n\lambda^{1/\alpha}}}M_1\sigma\sqrt{\log(\frac{8}{\delta})} \leq \frac{1}{2}$,

$$\left\|\boldsymbol{X}(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)\right\|_\infty \leq 2cM_1\left(\sigma\log(2/\delta')\frac{1}{\sqrt{n}}\lambda^{-1/4\alpha} + \lambda^{1/4}\,\|\boldsymbol{\theta}^*\|\right). \tag{E.10}$$

With $\lambda \gtrsim \left(\frac{\log n}{\sqrt{n}}\right)^{2\alpha} \frac{1}{g(n)^{\alpha/2}}$,

$$\frac{cC_1 \log n}{\sqrt{n\lambda^{1/\alpha}}} \lesssim cC_1 g(n)^{1/4} \to 0$$

so for n large enough, we do have $\frac{cC_1 M_1 \sigma \sqrt{\log(\frac{8}{\delta})}}{\sqrt{n\lambda^{1/\alpha}}} \leq \frac{1}{2}$.

Finally, we can plug back Eq. (E.10) into Eq. (E.9) with probability larger than $1-(n+1)\delta'$ conditionally on $\mathcal{B}$:

$$\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle\right| \leq c\left(\frac{\log n}{\sqrt{n\lambda^{1/\alpha}}} 2cC_1 M_1^3 \sigma \sqrt{\log(\frac{8}{\delta})}\left[\sigma \log(2/\delta')\frac{1}{\sqrt{n}}\lambda^{-1/4\alpha} + \lambda^{1/4}\left\|\boldsymbol{\theta}^*\right\|\right]\right.$$
$$\left. + \sigma \log(2/\delta')\frac{M_1}{\sqrt{n}}\lambda^{-1/4\alpha} + M_1\lambda^{1/4}\left\|\boldsymbol{\theta}^*\right\|\right),$$

i.e. there is a constant $c' > 0$ such that, for n large enough and $\delta' = \frac{\delta}{2(n+1)} \leq \frac{\delta}{6} \in (0,1)$:

$$\mathbb{P}\left(\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle\right| \leq c' M_1 \left(\sigma \log(n/\delta)\frac{1}{\sqrt{n}}\lambda^{-1/4\alpha} + \lambda^{1/4}\left\|\boldsymbol{\theta}^*\right\|\right) \,\middle|\, \mathcal{B}\right) \geq 1 - \frac{\delta}{2}, \quad \text{(E.11)}$$

and thus

$$\mathbb{P}\left(\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle\right| \leq c' M_1 \left(\sigma \log(n/\delta)\frac{1}{\sqrt{n}}\lambda^{-1/4\alpha} + \lambda^{1/4}\left\|\boldsymbol{\theta}^*\right\|\right)\right) \geq (1-\frac{\delta}{2})^2 \geq 1-\delta. \quad \text{(E.12)}$$

If we set

$$\lambda = c_2\left(\frac{M_1\sigma\sqrt{\log(8/\delta)}\log n}{\sqrt{n}}\right)^{2\alpha}\frac{1}{g(n)^{\alpha/2}} = c_2\left(M_1\sigma\sqrt{\log(8/\delta)}\right)^{2\alpha}(\log(n))^{\alpha(2+\beta/2)}\, n^{-\alpha},$$

the probabilistic bound in (E.12) becomes:

$$M_1\sigma\log(n/\delta)\frac{n^{-1/4}}{\log(n)^{1/2+\beta/8}}\left(M_1\sigma\sqrt{\log(8/\delta)}\right)^{-1/2} + \left(M_1\sigma\sqrt{\log(8/\delta)}\right)^{\alpha/2}\frac{\log(n)^{\alpha(1/2+\beta/8)}}{n^{\alpha/4}} M_1\left\|\boldsymbol{\theta}^*\right\|. \quad \text{(E.13)}$$

With $\alpha = 1$ and $\beta \in (0,8)$, it yields, with probability at least $1-\delta$

$$\left|\left\langle \Phi(\boldsymbol{x}_0), \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\rangle\right| \lesssim \sqrt{M_1^3\sigma}n^{-1/4}(\log(\frac{n}{\delta}))^{3/2} \lesssim \sqrt{M_1^3\sigma}n^{-1/4}(\log(\frac{n}{\delta}))^3.$$

(We only care about the n exponent, so we chose 3 as an upper bound to the true $3/2$ for the logarithm exponent to lighten the expression.)

□

F Proof of Appendix B.1's in-sample oracle inequality

Theorem B.1 (Oracle inequality)**.** *Let $\delta \in (0,1)$. Then there exists a universal constant $C_1 > 0$ such that under Assumption B.1, the following holds with probability at least $1-\delta$:*

$$\mathcal{R}_{in}(\hat{f}) \leq \inf_{\epsilon > 0}\left\{(1+\epsilon)\mathcal{R}_{in}(f_{j^*}) + \frac{C_1 K}{nk}\left(1+\frac{1}{\epsilon}\right)\xi(\tilde{\sigma}, b, \eta, \delta)\right\}, \quad \text{(B.6)}$$

with

$$\xi(\tilde{\sigma}, b, \eta, \delta) = K\tilde{\sigma}^2\log^2(\frac{m}{\delta}) + Kb^2 + \frac{L^2 n\eta^4}{k}.$$

Remark F.1. Under that event, we also have this intermediate inequality that holds with probability at least $1-\delta$:

$$\sqrt{\mathcal{R}_{\hat{j}}} \leq \sqrt{\mathcal{R}_{j^*}} + \frac{C_2\sqrt{K}}{\sqrt{nk}}\left(\sqrt{K}\tilde{\sigma}\log(\frac{m}{\delta}) + \sqrt{K}b + \frac{\sqrt{n}\eta^2}{\sqrt{k}}\right),$$

where $C_2 > 0$.

Proof. We start by introducing some notations for the rest of the proof. For any function f, let $\boldsymbol{f} = (f(\boldsymbol{x}_1), \cdots, f(\boldsymbol{x}_n))$ be the evaluation vector over the samples. We also denote for all j, $\mathcal{R}_j := \mathcal{R}_{in}(f_j)$ and $\widehat{\mathcal{R}}_j := \widehat{\mathcal{R}}_{in}(f_j)$.

We use the definition of the selected index $j^* = \operatorname{argmin}_{j\in[m]} \mathcal{R}_j$. We first need to establish a decomposition result between $\mathcal{R}$ and $\hat{\mathcal{R}}$:

Fact F.1. *Let f be any candidate model. Then*

$$\mathcal{R}_{in}(f) = \widehat{\mathcal{R}}_{in}(f) - \widehat{\mathcal{R}}_{in}(f^*) + \frac{1}{n}\sum_{i=1}^{n}(a'(\tilde{f}(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i)))(f(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)).$$

Proof. We simply apply the law of cosines of Bregman divergence since

$$\mathcal{R}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n} D_a(f(\boldsymbol{x}_i), f^*(\boldsymbol{x}_i)), \quad \widehat{\mathcal{R}}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n} D_a(f(\boldsymbol{x}_i), \tilde{f}(\boldsymbol{x}_i)).$$

□

Applying Fact F.1 to $\mathcal{R}_{\hat{j}}$, it follows that:

$$\begin{aligned}
\mathcal{R}_{\hat{j}} &= \widehat{\mathcal{R}}_{\hat{j}} - \widehat{\mathcal{R}}(f^*) + \frac{1}{n}\sum_{i=1}^{n}(a'(\tilde{f}(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i)))(f_{\hat{j}}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)) \\
&\leq \widehat{\mathcal{R}}_j - \widehat{\mathcal{R}}(f^*) + \frac{1}{n}\sum_{i=1}^{n}(a'(\tilde{f}(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i)))(f_{\hat{j}}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)), \quad \forall j \in [m] \\
&= \mathcal{R}_j + \frac{1}{n}\sum_{i=1}^{n}(a'(\tilde{f}(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i)))(f_{\hat{j}}(\boldsymbol{x}_i) - f_j(\boldsymbol{x}_i)),
\end{aligned}$$

where the first inequality stems from the minimality of $\widehat{\mathcal{R}}_{\hat{j}}$, and the last equality is obtained by re-applying Fact F.1 to $\mathcal{R}_j$.

By the Mean Value Theorem, we can write

$$a'(\tilde{f}(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i)) = a''(q_i)(\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i))$$

for some q_i between $f^*(\boldsymbol{x}_i)$ and $\tilde{f}(\boldsymbol{x}_i)$. Denote $\widetilde{\boldsymbol{D}}$ the positive diagonal matrix of entries $a''(q_i)$ that induces an inner product and a (weighted) norm so that the previous inequality becomes:

$$\mathcal{R}_{\hat{j}} \leq \mathcal{R}_j + \frac{1}{n}\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \boldsymbol{B} + \tilde{\boldsymbol{\varepsilon}} \right\rangle_{\widetilde{\boldsymbol{D}}}. \tag{F.1}$$

We cannot directly use a sub-exponential tail bound for the inner product because $\tilde{\boldsymbol{\varepsilon}}$ and $\widetilde{\boldsymbol{D}}$ are correlated. Thus we introduce the intermediate weight matrix $\overline{\boldsymbol{D}}$ with components $\overline{D}_{ii} = a''(f^*(\boldsymbol{x}_i))$ and use the fact that the perturbation introduced is small thanks to the consistency rate assumed.

In the remainder of this proof, we will use repeatedly the following inequalities linking the $\widetilde{\boldsymbol{D}}$-weighted norm, the $\overline{\boldsymbol{D}}$-weighted norm, and the standard Euclidean norm:

Fact F.2. *For all $z \in \mathbb{R}^n$:*

1. $\|z\|_2 \leq \frac{1}{\sqrt{k}} \|z\|_{\widetilde{\boldsymbol{D}}}$, $\quad \|z\|_{\widetilde{\boldsymbol{D}}} \leq \sqrt{K} \|z\|_2$.

2. $\|z\|_{\overline{\boldsymbol{D}}} \leq \sqrt{2} \|z\|_{\widetilde{\boldsymbol{D}}}$.

Proof. Since, we assume that the $\left\|\tilde{\boldsymbol{f}} - \boldsymbol{f}^*\right\|_\infty \leq \eta \leq C$, and $\|\boldsymbol{f}^*\|_\infty \leq C$, we can leverage our assumptions on the local properties of the log-partition function a in $[-2C, 2C]$. Part 1 directly follows from the definition of the weighted norm and boundedness of a''. For part 2, we obtain a stronger bound by leveraging the Lipschitzness assumption on a'' and the uniform consistency of the imputation model:

$$\left|\|z\|_{\overline{\boldsymbol{D}}}^2 - \|z\|_{\widetilde{\boldsymbol{D}}}^2\right| \leq \sum_i z_i^2 \left|a''(q_i) - a''(f^*(\boldsymbol{x}_i))\right| \leq \frac{L\eta}{k} \sum_i k z_i^2,$$

hence, via the strong convexity assumption,

$$\left|\|z\|_{\overline{\boldsymbol{D}}}^2 - \|z\|_{\widetilde{\boldsymbol{D}}}^2\right| \leq \frac{L\eta}{k} \sum_i a''(q_i) z_i^2 \leq \frac{L\eta}{k} \|z\|_{\widetilde{\boldsymbol{D}}}^2 .$$

Therefore, since $\frac{L\eta}{k} < 1$ according to the assumptions,

$$\|z\|_{\overline{\boldsymbol{D}}}^2 \leq 2 \|z\|_{\widetilde{\boldsymbol{D}}}^2 .$$

□

To make progress on inequality (F.1), we decompose the inner product into three terms that we can tackle independently:

$$\mathcal{R}_{\hat{j}} \leq \mathcal{R}_j + \frac{1}{n}\left[\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \boldsymbol{B}\right\rangle_{\widetilde{\boldsymbol{D}}} + \left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \tilde{\boldsymbol{\varepsilon}}\right\rangle_{\overline{\boldsymbol{D}}} + \left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \tilde{\boldsymbol{\varepsilon}}\right\rangle_{\widetilde{\boldsymbol{D}} - \overline{\boldsymbol{D}}}\right].$$

- Cauchy-Schwarz for the first term yields:

  $$\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \boldsymbol{B}\right\rangle_{\widetilde{\boldsymbol{D}}} \leq \left\|\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j\right\|_{\widetilde{\boldsymbol{D}}} \|\boldsymbol{B}\|_{\widetilde{\boldsymbol{D}}} \leq \sqrt{K} \left\|\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j\right\|_{\widetilde{\boldsymbol{D}}} \|\boldsymbol{B}\|_2 \leq \sqrt{K}\, b \left\|\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j\right\|_{\widetilde{\boldsymbol{D}}},$$

  where the last inequality holds on the event of item 3 in Assumption B.1.

- Cauchy-Schwarz for the third perturbation term gives:

  $$\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \tilde{\boldsymbol{\varepsilon}}\right\rangle_{\widetilde{\boldsymbol{D}} - \bar{\boldsymbol{D}}} \leq \left\|\widetilde{\boldsymbol{D}} - \bar{\boldsymbol{D}}\right\| \|\tilde{\boldsymbol{\varepsilon}}\|_2 \left\|\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j\right\|_2 \leq \frac{L}{\sqrt{k}} \max_{i\in[n]} \left|\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right| \cdot \|\tilde{\boldsymbol{\varepsilon}}\|_2 \cdot \left\|\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j\right\|_{\widetilde{\boldsymbol{D}}},$$

  where we used $\left\|\widetilde{\boldsymbol{D}} - \bar{\boldsymbol{D}}\right\| = \max_i \left|a''(q_i) - a''(f^*(\boldsymbol{x}_i))\right| \leq L \max_i \left|\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right|$. We will see that the consistency rate of our estimator is good enough to make this term of the same order as the two others.

  We can improve the $\max_{i\in[n]} \left|\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right| \cdot \|\tilde{\boldsymbol{\varepsilon}}\|_2$ factor by noting that:

  By assumption, $\forall i \in [n], |B_i + \tilde{\boldsymbol{\varepsilon}}_i| \leq \eta$, hence $|\tilde{\boldsymbol{\varepsilon}}_i| \leq \eta + |B_i|$ and $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2 \leq 2(n\eta^2 + \|\boldsymbol{B}\|_2^2) \leq 2(n\eta^2 + b^2)$, pathwise on the event of item 3 in Assumption B.1.

  Therefore,

  $$\max_{i\in[n]} \left|\tilde{f}(\boldsymbol{x}_i) - f^*(\boldsymbol{x}_i)\right|^2 \cdot \|\tilde{\boldsymbol{\varepsilon}}\|_2^2 \leq 2(n\eta^4 + \eta^2 b^2). \tag{F.2}$$

- For the second term, first bound its sub-exponential norm:

  $$\left\|\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \tilde{\boldsymbol{\varepsilon}}\right\rangle_{\overline{\boldsymbol{D}}}\right\|_{\psi_1} = \left\|\overline{\boldsymbol{D}}(\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j)\right\|_2 \left\|\left\langle \frac{\overline{\boldsymbol{D}}(\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j)}{\left\|\overline{\boldsymbol{D}}(\boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j)\right\|_2}, \tilde{\boldsymbol{\varepsilon}}\right\rangle_{\overline{\boldsymbol{D}}}\right\|_{\psi_1}$$

$$\leq \sqrt{K}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\overline{\boldsymbol{D}}}\left\|\left\langle\frac{\overline{\boldsymbol{D}}(\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j)}{\left\|\overline{\boldsymbol{D}}(\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j)\right\|_2},\tilde{\boldsymbol{\varepsilon}}\right\rangle_{\overline{\boldsymbol{D}}}\right\|_{\psi_1}$$
$$\leq \tilde{\sigma}\sqrt{K}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\overline{\boldsymbol{D}}}$$
$$\leq \tilde{\sigma}\sqrt{2K}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\tilde{\boldsymbol{D}}}.$$

Hence, the sub-exponential property and union bound gives that the event

$$\mathcal{A}=\left\{\forall j\in[m],\left\langle\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j,\tilde{\boldsymbol{\varepsilon}}\right\rangle_{\overline{\boldsymbol{D}}}\leq\sqrt{2K}\tilde{\sigma}\log(4m/\delta)\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\tilde{\boldsymbol{D}}}\right\}$$

holds with probability at least $1-\delta/2$. We place ourselves under $\mathcal{A}$.

So we obtain with probability at least $1-\delta$:

$$\mathcal{R}_{\hat{j}}\leq\mathcal{R}_j+\frac{2\xi}{n}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\tilde{\boldsymbol{D}}}, \tag{F.3}$$

where $\xi=(\sqrt{2K}\tilde{\sigma}\log(4m/\delta)+\sqrt{K}b+\frac{L\max_{i\in[n]}\left|\tilde{f}(\boldsymbol{x}_i)-f^*(\boldsymbol{x}_i)\right|\cdot\|\tilde{\boldsymbol{\epsilon}}\|_2}{\sqrt{k}})/2$.

Then, using part 1 of Fact F.2, the triangle inequality and Eq. (A.2), we get:

$$\frac{1}{\sqrt{n}}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_{\tilde{\boldsymbol{D}}}\leq\sqrt{\frac{K}{n}}\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}_j\right\|_2\leq\sqrt{\frac{K}{n}}\left(\left\|\boldsymbol{f}_{\hat{j}}-\boldsymbol{f}^*\right\|_2+\|\boldsymbol{f}_j-\boldsymbol{f}^*\|_2\right)\leq\sqrt{\frac{2K}{k}}\left(\sqrt{\mathcal{R}_j}+\sqrt{\mathcal{R}_{\hat{j}}}\right).$$

Hence, with $\tilde{\xi}=\sqrt{\frac{2K}{k}}\xi/\sqrt{n}$, we get:

$$\mathcal{R}_{\hat{j}}\leq\mathcal{R}_j+2\tilde{\xi}\left(\sqrt{\mathcal{R}_j}+\sqrt{\mathcal{R}_{\hat{j}}}\right),$$

and we can now complete the square:

- Rearranging:

  $$\mathcal{R}_{\hat{j}}-2\tilde{\xi}\sqrt{\mathcal{R}_{\hat{j}}}\leq\mathcal{R}_j+2\tilde{\xi}\sqrt{\mathcal{R}_j}\Leftrightarrow(\sqrt{\mathcal{R}_{\hat{j}}}-\tilde{\xi})^2-\tilde{\xi}^2\leq(\sqrt{\mathcal{R}_j}+\tilde{\xi})^2-\tilde{\xi}^2.$$

  Thus,

  $$\sqrt{\mathcal{R}_{\hat{j}}}\leq\sqrt{\mathcal{R}_j}+2\tilde{\xi}.$$

  Since this is valid for all j, an intermediate result is thus:

  $$\sqrt{\mathcal{R}_{\hat{j}}}\leq\sqrt{\mathcal{R}_{j^*}}+2\tilde{\xi}.$$

- Squaring on both sides, we obtain:

  $$\mathcal{R}_{\hat{j}}\leq\mathcal{R}_{j^*}+2\sqrt{4\tilde{\xi}^2\mathcal{R}_{j^*}}+4\tilde{\xi}^2.$$

  Using the simple inequality $2ab\leq a^2+b^2$ and the "Rob Peter to pay Paul" trick with any $\epsilon>0$, i.e. $2ab=2\frac{a}{\epsilon}\epsilon b\leq\frac{a^2}{\epsilon^2}+\epsilon^2b^2$, we get:

  $$\mathcal{R}_{\hat{j}}\leq(1+\epsilon)\mathcal{R}_{j^*}+4\tilde{\xi}^2(1+\frac{1}{\epsilon}).$$

  Finally, to get the final formula of the theorem, we leverage Eq. (F.2) to bound $\tilde{\xi}^2$ as:

  $$\tilde{\xi}^2\lesssim\frac{K}{kn}(K\tilde{\sigma}^2\log^2(m/\delta)+b^2(K+2L^2\eta^2/k)+\frac{L^2n\eta^4}{k}).$$

  Using $\eta\leq\sqrt{kK}/L$ and taking the infimum on ϵ yields the final result.

□

Remark F.2 (The argument is pathwise in $\boldsymbol{B}$)**.** The bias $\boldsymbol{B}$ enters the proof above only through the Cauchy–Schwarz bound on $\left\langle \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j, \boldsymbol{B} \right\rangle_{\widetilde{\boldsymbol{D}}}$ and the bound $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2 \le 2(n\eta^2 + \|\boldsymbol{B}\|_2^2)$, which hold for every realization. Neither the law of $\boldsymbol{B}$ nor its dependence on $\tilde{\boldsymbol{\varepsilon}}$ or on $\hat{j}$ plays any role; only $\|\boldsymbol{B}\|_2 \le b$ is used. Concentration is only needed for $\tilde{\boldsymbol{\varepsilon}}$, through the event $\mathcal{A}$, whose union bound over $j \in [m]$ covers the data-dependent $\hat{j}$.

We deduce a bound on the expectation:

Corollary F.1. *Under the same assumptions as Theorem B.1, and assuming in addition that the bounds of item 3 in Assumption B.1 hold almost surely, we deduce:*

$$\mathbb{E}\left[\mathcal{R}_{\hat{j}}\right] \le \inf_{\epsilon>0}\left\{(1+\epsilon)\mathcal{R}_{j^*} + \frac{C_1 K}{nk}\left(1+\frac{1}{\epsilon}\right)\left(Kb^2 + \frac{L^2 n\eta^4}{k} + K\tilde{\sigma}^2(1+(\log(m)+1)^2)\right)\right\}.$$

Proof. Denoting $W = [\mathcal{R}_{\hat{j}} - (1+\epsilon)\mathcal{R}_{j^*} - \frac{C_1K}{nk}(1+\frac{1}{\epsilon})(Kb^2 + L^2n\eta^4/k)]/\frac{C_1K}{nk}(1+\frac{1}{\epsilon})K\tilde{\sigma}^2$, we have from the theorem:

$$\mathbb{P}\left(W \ge \log^2(m/\delta)\right) \le \delta, \forall \delta \in (0,1).$$

Let $t = \log^2(m/\delta)$ such that:

$$\mathbb{P}\left(W \ge t\right) \le m e^{-\sqrt{t}}, \forall t \ge \log^2(m).$$

Integrating to get the expectation yields:

$$\begin{aligned}\mathbb{E}\left[W\right] = \int_0^\infty \mathbb{P}\left(W > t\right) dt &\le \int_0^{\log^2(m)} 1 dt + \int_{\log^2(m)}^\infty m e^{-\sqrt{t}} dt \\ &= \log^2(m) + 2(\log(m)+1) = (1+\log(m))^2 + 1,\end{aligned}$$

since $\int_{u^2}^\infty e^{-\sqrt{t}} dt = 2(u+1)e^{-u}$. Rearranging the terms and taking the infimum gives the desired result. □

G Useful lemmas

G.1 Lemmas for the GLM estimator

Lemma G.1. *Let the kernel regularity Assumption 4.1 hold, with spectral eigendecay of the form* $\mu_j \le A j^{-2\alpha}, \alpha \ge 1$*, and let* $\lambda \in (0,1]$*. Then the effective dimension* $r(\lambda) := \sum_{j=1}^\infty \frac{\mu_j}{\mu_j + \lambda}$ *satisfies for a universal constant* $C > 0$*:*

$$r(\lambda) \le C\lambda^{-1/2\alpha}.$$

In particular, if the trace class operator $\boldsymbol{\Sigma}$ *has eigenvalues* μ_j*, then* $\mathrm{Tr}((\boldsymbol{\Sigma}+\lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}) = r(\lambda) \lesssim \lambda^{-1/2\alpha}$.

Likewise, for a vector $\Phi(\boldsymbol{x})$ *of the RKHS with basis decomposition* $\Phi(\boldsymbol{x}) = (\sqrt{\mu_1}\phi_1(\boldsymbol{x}), \sqrt{\mu_2}\phi_2(\boldsymbol{x})...)$*, where the basis functions are uniformly bounded by* M_1*, the quadratic form* $\left\langle \Phi(\boldsymbol{x}), (\boldsymbol{\Sigma}+\lambda\boldsymbol{I})^{-1}\Phi(\boldsymbol{x})\right\rangle \le M_1^2 r(\lambda) \lesssim M_1^2 \lambda^{-1/2\alpha}$.

Proof. Split the sum before and after $j = \lambda^{-1/2\alpha}$ and use the asymptotic tail of the series. Without loss of generality, we assume $A = 1$ (for $A \ne 1$, use $\lambda = \lambda/A$ in the formula).

$$\begin{aligned}\sum_{j=1}^\infty \frac{\mu_j}{\mu_j+\lambda} &= \sum_{j \le \lambda^{-1/2\alpha}} \frac{1}{1+\lambda j^{2\alpha}} + \sum_{j > \lambda^{-1/2\alpha}} \frac{1}{1+\lambda j^{2\alpha}} \\ &\le \sum_{j \le \lambda^{-1/2\alpha}} 1 + \sum_{j > \lambda^{-1/2\alpha}} \frac{1}{\lambda j^{2\alpha}}\end{aligned}$$

$$\leq \lambda^{-1/2\alpha} + \frac{1}{\lambda(2\alpha-1)\lambda^{-(2\alpha-1)/2\alpha}}$$
$$\leq \frac{2\alpha}{2\alpha-1}\lambda^{-1/2\alpha}.$$

We find that $C = \frac{2\alpha A^{1/2\alpha}}{2\alpha-1}$ in general. □

Remark G.1. By the eigendecay for the source distribution, applying Lemma G.1 with $\alpha = 1$ directly gives for $\lambda \in (0, 1)$

$$\mathrm{Tr}(\boldsymbol{C}_\lambda) \lesssim \lambda^{-1/2}.$$

G.2 Technical lemmas

Lemma G.2. *Let $\boldsymbol{x}$ be a centered sub-exponential vector with $\|x_i\|_{\psi_1} \leq 1, \forall i \in [n]$, $\boldsymbol{A} \succeq 0$ a fixed PSD matrix in $\mathbb{R}^{n\times n}$. Then there exists a universal constant $C > 0$, such that $\forall \delta \in (0, \frac{2}{e}]$, with probability at least $1 - \delta$,*

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \leq C\,\mathrm{Tr}(\boldsymbol{A})\log^2(n)\log(\frac{2}{\delta})$$

when $n \geq \left(\frac{4}{\delta}\right)^{1/3}$.

Proof. We will truncate the sub-exponential variables to get sub-Gaussian variables and bound the gap between the quadratic forms. Then use lemma D.1 in (Wang, 2026).

- First by definition of sub-exponential,

  $$\forall i, \mathbb{P}\left(|x_i| > C\log n\right) \leq 2n^{-C}$$

  for any constant $C > 0$. Hence if we define the truncated version $\bar{\boldsymbol{x}}$ of components $\bar{x}_i := x_i \mathbb{1}(|x_i| \leq C\log n)$, by union bound, $\boldsymbol{x} = \bar{\boldsymbol{x}}$ with probability larger than $1 - 2n^{1-C}$. Denote this event $\mathcal{A}$. Also, by definition of $\bar{\boldsymbol{x}}$, we have immediately that $\|\bar{\boldsymbol{x}}\| \leq C\sqrt{n}\log n$ almost surely.

- $\bar{\boldsymbol{x}}$ is not centered anymore so let its mean be $\boldsymbol{\mu}$. Then $\forall i, |\mu_i| = \mathbb{E}\left[x_i \mathbb{1}(|x_i| \leq C\log n)\right] = \mathbb{E}\left[x_i \mathbb{1}(|x_i| > C\log n)\right]$ since $\mathbb{E}\left[x_i\right] = 0$. So Cauchy-Schwarz yields:

  $$|\mu_i| \leq \sqrt{\mathbb{E}\left[|x_i|^2\right]\mathbb{P}\left(|x_i| > C\log n\right)} \leq 2\sqrt{2}n^{-C/2} \lesssim n^{-C/2},$$

  where the first factor is bounded by a constant by the sub-exponential property, and the second obeys the tail bound mentioned before. Hence:

  $$\|\boldsymbol{\mu}\| \lesssim \sqrt{n}n^{-C/2} = n^{(1-C)/2}.$$

- Next define the centered version $\tilde{\boldsymbol{x}} := \bar{\boldsymbol{x}} - \boldsymbol{\mu}$. It is centered and bounded almost surely: $\|\tilde{\boldsymbol{x}}\|_\infty \leq 2C\log n$ hence it is sub-Gaussian with $\|\tilde{\boldsymbol{x}}\|_{\psi_2} \lesssim C\log n$. Thus its quadratic form satisfies a tail bound as proposed in lemma D.1 of (Wang, 2026) for a universal constant $C' > 0, \forall \delta \in (0, \frac{2}{e}]$, with probability at least $1 - \frac{\delta}{2}$:

  $$\tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}} \leq C' \|\tilde{\boldsymbol{x}}\|_{\psi_2}^2 \mathrm{Tr}(\boldsymbol{A})\log(2/\delta) \lesssim \log^2(n)\,\mathrm{Tr}(\boldsymbol{A})\log(2/\delta).$$

  Denote this event $\mathcal{B}$.

- Finally, we can bound the difference $\left|\tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{x}}\right| = \left|\boldsymbol{\mu}^\top \boldsymbol{A}(2\bar{\boldsymbol{x}} - \boldsymbol{\mu})\right| \leq \|\boldsymbol{A}\|\left(2\|\boldsymbol{\mu}\|\,\|\boldsymbol{x}\| + \|\boldsymbol{\mu}\|^2\right) \leq \|\boldsymbol{A}\|\left(n^{1-C} + n^{1-C/2}\log n\right)$.

Putting everything together, under the event $\mathcal{A} \cap \mathcal{B}$ which holds with probability larger than $1 - 2n^{1-C} - \frac{\delta}{2}$ by union bound, we get: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}} + (\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \bar{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}}) \lesssim \|\boldsymbol{A}\|\left(n^{1-C} + n^{1-C/2}\log n\right) + \log^2 n\,\mathrm{Tr}(\boldsymbol{A})\log(2/\delta)$.

Choosing $C = 4$, since $\log(2/\delta) \geq 1$ for $\delta \leq 2/e$, and $\|\boldsymbol{A}\| \leq \operatorname{Tr}(\boldsymbol{A})$, yields:

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \lesssim \operatorname{Tr}(\boldsymbol{A}) \log(\frac{2}{\delta}) \left(\log^2(n) + \frac{1}{n^3} + \frac{\log(n)}{n} \right)$$

with probability larger than $1 - \frac{2}{n^3} - \frac{\delta}{2}$. For $n \geq \left(\frac{4}{\delta}\right)^{1/3}$, the probability is larger than $1 - \delta$ as wanted. For $\delta = 0.05$, this threshold is just $n \geq 5$. Removing the negligible terms yields the desired result.

□

Lemma G.3. *Let Assumptions 2.1 and 4.1 hold, with M_2 the bound on the covariates $\Phi(\boldsymbol{x}_i)$. Choose $\gamma \in (0, 1)$, $\delta \in (0, 1/e]$, and $\lambda \geq \frac{4M_2^2 \log(14n/\delta)}{\gamma^2 n}$. Then the following event holds with probability at least $1 - \delta$:*

$$\mathbb{P}\left((1-\gamma)(\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \preceq \hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I} \preceq (1+\gamma)(\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \right) \geq 1 - \delta.$$

Under that event, the following inequalities are true:

1. *For any vector $\boldsymbol{a} \in \mathbb{H}$, $\left\langle \boldsymbol{a}, (\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \boldsymbol{a} \right\rangle \leq (1-\gamma)^{-1} \left\langle \boldsymbol{a}, (\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{a} \right\rangle$.*
2. *$\operatorname{Tr}((\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}}) \leq M_1^2 (1-\gamma)^{-1} \operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma})$, where $M_1 > 0$ is the uniform bound on the eigenfunctions.*

Proof. The first part is just a repetition of corollary D.1 in (Wang, 2026). Then, the first inequality is a direct corollary from the main sandwich inequality, by definition of the PSD order. Finally, for the second inequality, observe that the sandwich inequality implies:

$$\operatorname{Tr}((\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}}) \leq \frac{1}{1-\gamma} \operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}}).$$

Define $\tilde{\Phi}_i := (\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2} \Phi(\boldsymbol{x}_i)$ such that

$$\operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}}) = \operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2} \hat{\boldsymbol{\Sigma}} (\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} \left\| \tilde{\Phi}_i \right\|^2.$$

Since by uniform boundedness of the eigenfunctions,

$$\left\| \tilde{\Phi}_i \right\|^2 = \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} |\phi_j(\boldsymbol{x}_i)|^2 \leq M_1^2 \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} = M_1^2 \operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma}),$$

we get:

$$\operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \hat{\boldsymbol{\Sigma}}) \leq M_1^2 \operatorname{Tr}((\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Sigma}),$$

and the result follows.

□

G.3 Relating the source covariance operator to the target's via the effective sample size

Through the effective dimension defined as $n_{\text{eff}} = \sup \left\{ t \leq n | t\boldsymbol{\Sigma}_0 \preceq n\boldsymbol{\Sigma} + \frac{\mu^2}{k} \boldsymbol{I} \right\}$ in equation Eq. (4.1), we can relate expressions depending on $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}_0$-only expressions. Take $\lambda \geq \frac{\mu^2}{n}$, and let $r = \frac{n_{\text{eff}}}{n}$. On the one hand we have

$$2(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I}) \succeq (k\boldsymbol{\Sigma} + \mu^2/n\boldsymbol{I}) + \lambda \boldsymbol{I} \succeq rk\boldsymbol{\Sigma}_0 + \lambda \boldsymbol{I} = r(k\boldsymbol{\Sigma}_0 + r^{-1}\lambda \boldsymbol{I}),$$

hence

$$(k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1} \preceq 2r^{-1}(k\boldsymbol{\Sigma}_0 + r^{-1}\lambda \boldsymbol{I})^{-1}, \tag{G.1}$$

and

$$\operatorname{Tr}(\boldsymbol{S}_\lambda) \leq 2r^{-1} \operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \lambda r^{-1}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0). \tag{G.2}$$

On the other hand, we have simply:

$$k\boldsymbol{\Sigma} + \lambda \boldsymbol{I} \succeq rk\boldsymbol{\Sigma}_0,$$

hence:

$$\|\boldsymbol{S}_\lambda\| \leq r^{-1}k^{-1}. \tag{G.3}$$

H Complement to the numerical experiments from Section 5 and ablation study

H.1 Implementation details

We optimize the ridge-regularized kernel GLM Eq. (2.7) in function space $\mathbf{f} = \boldsymbol{K}\boldsymbol{\alpha}$ via Fisher scoring (McCullagh and Nelder, 1989), i.e. Newton's method with the Hessian replaced by the expected Hessian (Fisher information). At each iteration, with current evaluations $\eta_i = f(\boldsymbol{x}_i)$, mean $\mu_i = \mathbb{E}[y_i \mid \eta_i]$, and variance $V_i = \operatorname{Var}[y_i \mid \eta_i]$, we compute the standard GLM weights $\boldsymbol{W} = \operatorname{diag}((\frac{d\mu_i}{d\eta_i})^2/V_i)$ and pseudo-responses $z_i = \eta_i + (y_i - \mu_i)/\frac{d\mu_i}{d\eta_i}$. The resulting iteratively reweighted least squares (IRLS) update requires solving $(\boldsymbol{K}\boldsymbol{W} + n\lambda\boldsymbol{I})\mathbf{f} = \boldsymbol{K}\boldsymbol{W}\boldsymbol{z}$. $\boldsymbol{K} \in \mathbb{R}^{n\times n}$ is the kernel matrix introduced in equation Eq. (2.8).

To ensure stable convergence of our conjugate gradient (Hestenes and Stiefel, 1952) (CG) solver, we avoid extreme weight scalings by defining $\boldsymbol{S} = \boldsymbol{W}^{1/2}$ and substituting $\boldsymbol{u} = \boldsymbol{S}\mathbf{f}$ to solve the equivalent symmetric positive definite system:

$$(\boldsymbol{S}\boldsymbol{K}\boldsymbol{S} + n\lambda\boldsymbol{I})\boldsymbol{u} = \boldsymbol{S}\boldsymbol{K}\boldsymbol{W}\boldsymbol{z}$$

We then recover $\mathbf{f} = \boldsymbol{S}^{-1}\boldsymbol{u}$. After convergence, the kernel coefficients needed to evaluate $\hat{f} = \sum_i \alpha_i K(\boldsymbol{x}_i, \cdot)$ at new (e.g., target) points are obtained without solving another linear system: rearranging $(\boldsymbol{K}\boldsymbol{W} + n\lambda\boldsymbol{I})\mathbf{f} = \boldsymbol{K}\boldsymbol{W}\boldsymbol{z}$ as $\mathbf{f} = \boldsymbol{K}\,\boldsymbol{W}(\boldsymbol{z} - \mathbf{f})/(n\lambda)$ gives $\boldsymbol{\alpha} = \frac{1}{n\lambda}\boldsymbol{W}(\boldsymbol{z} - \mathbf{f})$.

For large-scale problems, forming the dense $n \times n$ matrix $\boldsymbol{K}$ introduces an $\mathcal{O}(n^2)$ memory bottleneck. We circumvent this by implementing the inner CG solver using KeOps (Charlier et al., 2021) for on-the-fly GPU kernel-vector products. This strategy reduces the memory footprint to $\mathcal{O}(nd)$ without altering the exact IRLS update equations. Full algorithmic details and solvers are available at `https://github.com/nathanweill/pseudolabel-uda-experiments`.

H.2 Additional details and results for synthetic data experiment

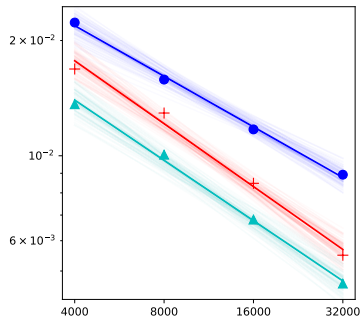
Kernel logistic regression specifics. In the following experiments, we focus on kernel logistic regression, for which the log-partition function is $a(u) = \log(1 + e^u)$, and its first derivative is the sigmoid function $a'(u) = \frac{e^u}{1+e^u} = \sigma(u)$ that maps real values to $[0, 1]$. If the true underlying function is f^*, then $y_i \mid \boldsymbol{x}_i \sim \text{Bernoulli}(\sigma(f^*(\boldsymbol{x}_i)))$. We can thus denote the conditional mean as a probability $p_i^* := \sigma(f^*(\boldsymbol{x}_i))$. Equivalently, $f^*(\boldsymbol{x}_i) = \log(\frac{p_i^*}{1-p_i^*}) = \text{logit}(p_i^*)$. As such, for any estimator f of f^*, if we denote for each sample $p_i = \sigma(f(\boldsymbol{x}_i))$, the in-sample version of the GLM risk Eq. (2.5) recovers the usual log-loss with "soft labels" or cross-entropy between the true distribution p^* and the estimate p:

$$R_{in}(f) = \frac{1}{n}\sum_{i=1}^{n}\left[a(f(\boldsymbol{x}_i)) - a'(f^*(\boldsymbol{x}_i))f(\boldsymbol{x}_i)\right] = \frac{1}{n}\sum_{i=1}^{n}\left[-(1-p_i^*)\log(1-p_i) - p_i^*\log(p_i)\right].$$

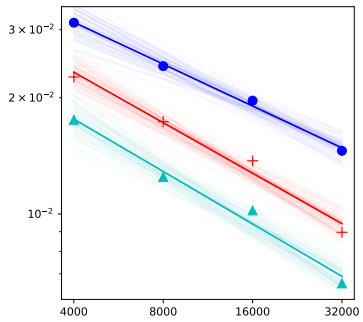
Figure 1b shows the results for the same experiments as in Section 5.1, with covariate shift strength $B = n^{0.45}$. The fitted exponents for all shift levels $B = n^\gamma$ considered in Section 5.1 are reported in Table 4; the rows $\gamma = 0.40$ and $\gamma = 0.45$ correspond to Figures 1a and 1b.

Table 4: Fitted decay exponents κ of the target excess risk (bootstrap SE) for $B = n^{\gamma}$, i.e., $n_{\text{eff}} \asymp n^{1-\gamma}$.

γ	Minimax $2(1-\gamma)/3$	Pseudo-labeling	Oracle	Naive
0.40	0.400	0.546 (0.047)	0.523 (0.051)	0.439 (0.043)
0.45	0.367	0.434 (0.049)	0.451 (0.053)	0.360 (0.047)
0.50	0.333	0.386 (0.059)	0.369 (0.058)	0.285 (0.048)
0.55	0.300	0.402 (0.059)	0.342 (0.057)	0.231 (0.048)
0.60	0.267	0.362 (0.060)	0.273 (0.059)	0.111 (0.046)
0.65	0.233	0.301 (0.060)	0.176 (0.064)	0.008 (0.042)



(a) $B = n^{0.4}$



(b) $B = n^{0.45}$

Figure 1: Comparisons of the three approaches on a log-log plot. x-axis: n. y-axis: excess risk. Red crosses: pseudo-labeling. Cyan triangles: the oracle method. Blue circles: the naive method. Solid lines: fitted power laws; thin shaded regions: linear fits of 100 bootstrap replicates.

H.3 Ablation study: validating the undersmoothed imputation penalty

Our theoretical analysis dictates that the imputation model must be undersmoothed to ensure valid model selection with pseudo-labels. To empirically validate this insight, we ran an ablation study on the imputation model's penalty ($\tilde{\lambda}$). While our theoretical guarantees (Theorem B.1) hold for finite samples, isolating the leading-order effects that govern the optimal penalty scaling requires a sufficiently large sample size. We evaluate our method using $n = 16,000$ to empirically validate this regime without interference from finite-sample noise. To ensure statistical significance, we sweep across a grid of penalty values and measure the empirical target excess risk of the final estimator over 50 independent Monte Carlo trials. We use the same setup as the previous section, i.e. same feature space, response model, and kernel, with covariate shift strength $B = n^{0.45}$. Given the source-trained candidate models, we select the target-optimal one using target pseudo-labels obtained with an imputation model. The only varying parameter is the imputation model's penalty. Once selected, we report the final estimator's target excess risk.

Figure 2 plots the pseudo-labeling final estimator's excess risk as a function of the imputer's penalty. The empirically optimal imputer's penalty matches our theoretical requirement for an undersmoothed imputation model, and yields a tenfold reduction in excess risk compared to a standard baseline penalty. Moreover, the improvement of the undersmoothed estimator over the baseline is highly statistically significant across all random seeds.

Sensitivity to $\tilde{\lambda}$. Within the undersmoothed regime, the final estimator is insensitive to $\tilde{\lambda}$: its target excess risk is 0.012 ± 0.001 at $\tilde{\lambda} = 10^{-6}$, 0.011 ± 0.001 at the empirical optimum 10^{-5}, and still 0.020 ± 0.002 at 10^{-4}, whereas the oversmoothed baseline 10^{-3} gives 0.114 ± 0.003. What matters is thus to avoid oversmoothing, which the rule $\tilde{\lambda} = \min \Lambda$ of Section 3 achieves without target labels. This rule is motivated by Theorem 4.1, where $\min \Lambda = \mu^2/n$ and $\tilde{\lambda} = \mu^2 \log^7(n) \log(n_0/\delta)/n$ share the unknown scale $\mu^2 = M_1^6 \sigma^2$, so that their ratio is a known logarithmic factor, which we drop. It remains a heuristic, since in practice Λ is built from source data rather than from the theoretical grid,

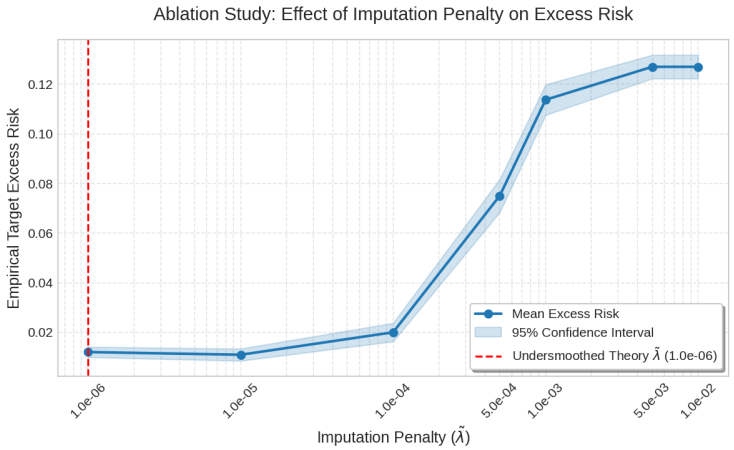


Figure 2: Pseudo-labeling final estimator's excess risk as a function of the imputer's penalty ($n = 16{,}000$, 50 trials). The rule $\tilde{\lambda} = \min \Lambda = 1/(10n) \approx 6.3 \times 10^{-6}$ used in Section 5.1 lies between the two leftmost points, in the flat region.

which depends on unknown constants. At small sample sizes, the variance of the imputer also matters, which motivates the repeated folds of Algorithm 2 for Raisin. Finally, undersmoothing cannot correct the part of the imputer's error due to a change in the conditional law, which lies outside our theory.

Again, full details and reproducible code can be found at `https://github.com/nathanweill/pseudolabel-uda-experiments`.

H.4 Details about the real data experiments

To first illustrate the effect of this shift, we train a ridge-regularized logistic regression baseline only on source data, with a grid of ridge regularization parameters λ. We keep 20% of the ID data and all of the OOD data for evaluation.

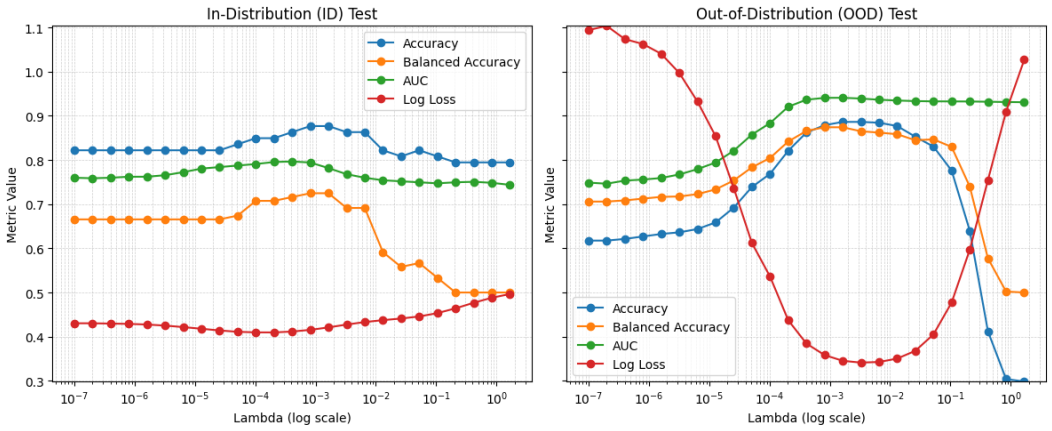


Figure 3: Baseline logistic regression performance on ID and OOD test data, when trained only on ID data, as a function of the ridge regularization parameter λ.

Figure 3 plots the results for four different metrics. We can observe that only a specific range of regularization (between 10^{-3} and 10^{-2}) can achieve good performance on the target data, different from the optimal one for the source. The goal of our method is to find a near-optimal value without seeing the target labels used to plot these curves.

Table 5: Risk mean estimate across seeds with confidence interval (CI).

Method	Mean	95% CI
Naive	0.442	[0.414, 0.469]
KMM + IW-CV	0.449	[0.431, 0.468]
KLIEP + IW-CV	0.434	[0.412, 0.465]
KRR pseudo-labeling (Wang, 2026)	0.406	[0.388, 0.424]
Pseudo-labeling	0.383	[0.372, 0.394]
Oracle	0.376	[0.362, 0.389]

Let us explain in more detail the experiments of Section 5.2. After sub-sampling, we have labeled in-distribution (ID) data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and labeled out-of-distribution (OOD) data $\mathcal{D}_0 = \{(x_{0,i}, y_{0,i})\}_{i=1}^{n_0}$ (OOD labels are used *only* for oracle benchmarking and final evaluation). We first split the OOD sample into a selection set and a test set, $\mathcal{D}_0 = \mathcal{D}_{0,\mathrm{sel}} \cup \mathcal{D}_{0,\mathrm{test}}$, and report all final metrics on $\mathcal{D}_{0,\mathrm{test}}$.

Recall that for a score function $f : \mathcal{X} \to \mathbb{R}$, the empirical logistic risk on a finite set S with targets $t \in [0, 1]$ is

$$R_S(f; t) \;:=\; \frac{1}{|S|} \sum_{x \in S} \Big(a(f(x)) - t(x)\, f(x) \Big), \qquad a(u) = \log(1 + e^u),$$

where $t(x) \in \{0, 1\}$ yields standard log-loss and $t(x) \in [0, 1]$ yields the corresponding *soft-label* risk.

As mentioned in Section 5.2, due to the small sample size, we use a slightly modified selection procedure compared to Algorithm 1 and Section 5.1, to deflate the variance of the final estimator. The chosen cross-validation procedure is designed so as to minimize the bias of the imputation model, while keeping its variance reasonable; in other words, it is intended to minimize its pointwise error. It is detailed in Algorithm 2. Each fitting in steps 2 and 5 is a ridge logistic regression task.

Refitting on all labeled ID data. Step 5 of Algorithm 2 refits the final model at the selected penalty on all labeled ID data, as is standard after cross-validation (James et al., 2021; Friedman et al., 2010). This is purely an empirical enhancement: our guarantees cover the model trained on a single split (Algorithm 1), and we do not prove that refitting preserves them. We use it only for Raisin, where a single split of the 900 instances leaves too little data for a stable final fit, and apply it identically to all methods, so that it cannot create the reported gaps.

Hyperparameters of the Raisin experiment. In the specific case of the Raisin dataset, we chose $\lambda_1 = 10^{-4}$ to $\lambda_m = 10^2$ for the candidates grid, and $\tilde{\lambda} = 10^{-4}$ for the imputer, i.e., $\tilde{\lambda} = \min \Lambda$. Then we ran Algorithm 2 over 100 random seeds. In Section 5.2, we reported the results for $K = 2, R = 6$. Table 5 provides the results for the same experiment, but using $K = 5, R = 2$ instead. In this setting, we also compare with the KRR pseudo-labeling method of Wang (2026), i.e., the same pipeline with square-loss candidates and imputer, evaluated with the same target log-loss. It improves on naive validation and on the density-ratio methods, but remains behind our GLM version (0.406 versus 0.383): even with oracle selection, its square-loss candidates only reach 0.443, reflecting their poor calibration for classification. The CI of this row is computed from its standard error.

Finally for this 5-fold (2 repeats) cross-validation experiment, we plotted the risk curves used for model selection for each of the three methods, described in step 3's equation (H.1), averaged over seeds in Figure 4. The pseudo-labeling method constructs a risk curve consistent with the oracle, thus achieving good model selection for the target. In contrast, the naive risk curve stays flat.

H.5 WILDS benchmark evaluation

Dataset and domain shift. Camelyon17 (Bandi et al., 2018) consists of histopathology patch images for tumor detection, collected across five hospitals using different scanners and staining protocols, as depicted in Figure 5. Following the standard WILDS (Koh et al., 2021) splits and the

Algorithm 2 Repeated K-Fold Cross-Validated Pseudo-Labeling (Practical Variant)

Input: Labeled ID data $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, labeled OOD data $\mathcal{D}_0 = \{(x_{0i}, y_{0i})\}_{i=1}^{n_0}$ split into $\mathcal{D}_{0,\text{sel}}$ and $\mathcal{D}_{0,\text{test}}$. OOD labels are *not* used by the pseudo-labeling method; they serve only for the oracle benchmark and final evaluation. Candidate grid $\Lambda = \{\lambda_1, \ldots, \lambda_m\}$, imputer penalty $\tilde{\lambda}$, number of folds K, number of repeats R.

Step 1 (Repeated K-fold partition). For each repeat $r \in [R]$, draw a fresh stratified (i.e. preserving class proportions) K-fold partition of the ID indices:

$$\{1, \ldots, n\} = \bigsqcup_{k=1}^{K} F_k^{(r)}, \qquad F_k^{(r)} \cap F_{k'}^{(r)} = \emptyset.$$

Define the candidate and imputer index sets for each fold (r, k):

$$I_{\text{cand}}^{(r,k)} := F_k^{(r)}, \qquad I_{\text{imp}}^{(r,k)} := \{1, \ldots, n\} \setminus F_k^{(r)}.$$

Step 2 (Per-fold training and evaluation). For each fold (r, k):

(a) **Imputer.** Fit $\tilde{f}^{(r,k)}$ on $\{(x_i, y_i)\}_{i \in I_{\text{imp}}^{(r,k)}}$ with penalty $\tilde{\lambda}$. Compute soft pseudo-labels:

$$p^{(r,k)}(x) := \sigma(\tilde{f}^{(r,k)}(x)), \quad x \in \mathcal{D}_{0,\text{sel}}, \quad \sigma(u) = (1 + e^{-u})^{-1}.$$

(b) **Candidates.** For each $\lambda_j \in \Lambda$, fit $f_{\lambda_j}^{(r,k)}$ on $\{(x_i, y_i)\}_{i \in I_{\text{cand}}^{(r,k)}}$.

(c) **Risk curves.** For each λ_j, compute:

$$\begin{aligned}
\widehat{R}_{\text{naive}}^{(r,k)}(\lambda_j) &:= R_{\{x_i : i \in I_{\text{imp}}^{(r,k)}\}}\big(f_{\lambda_j}^{(r,k)};\ y\big), \\
\widehat{R}_{\text{pseudo}}^{(r,k)}(\lambda_j) &:= R_{\mathcal{D}_{0,\text{sel}}}\big(f_{\lambda_j}^{(r,k)};\ p^{(r,k)}\big), \\
\widehat{R}_{\text{oracle}}^{(r,k)}(\lambda_j) &:= R_{\mathcal{D}_{0,\text{sel}}}\big(f_{\lambda_j}^{(r,k)};\ y_0\big).
\end{aligned}$$

Step 3 (Aggregation). Average over all folds and repeats:

$$\widehat{R}_\star(\lambda_j) := \frac{1}{KR} \sum_{r=1}^{R} \sum_{k=1}^{K} \widehat{R}_\star^{(r,k)}(\lambda_j), \qquad \star \in \{\text{naive, pseudo, oracle}\}. \tag{H.1}$$

Step 4 (Selection). Select regularization parameters independently:

$$\hat{\lambda}_\star \in \arg\min_{\lambda \in \Lambda} \widehat{R}_\star(\lambda), \qquad \star \in \{\text{naive, pseudo, oracle}\}.$$

Step 5 (Refit). For each $\star \in \{\text{naive, pseudo, oracle}\}$, refit a final model $\hat{f}_\star$ on *all* labeled ID data $\mathcal{D}$.

Step 6 (Evaluation). Evaluate each final model on the held-out OOD test set:

$$R_{\text{test}}(\hat{f}_\star) := R_{\mathcal{D}_{0,\text{test}}}\big(\hat{f}_\star;\ y_0\big).$$

Output: Test risks $R_{\text{test}}(\hat{f}_{\text{naive}})$, $R_{\text{test}}(\hat{f}_{\text{pseudo}})$, $R_{\text{test}}(\hat{f}_{\text{oracle}})$.

hospital labels of Figure 5, we use hospitals 1–3 as the labeled source domain and hospital 5 as the unlabeled target domain; hospital 4 (the WILDS out-of-distribution validation split) is not used. The source contains 302,436 training patches and 33,560 in-distribution validation patches, and the target contains 85,054 patches. The domain shift is structural — different tissue preparation and imaging conditions — and persists even with strong pretrained representations, making it a challenging and realistic UDA benchmark.

Model and features. We fine-tune a DINOv2 ViT-B/14 backbone (Oquab et al., 2023) pretrained on LVD-142M, with a linear classification head, end-to-end on source data using AdamW (weight decay 0.01, cosine annealing schedule, batch size 256). The epoch number t serves as the regularization parameter: early epochs correspond to strong regularization (small deviation from pretrained weights),

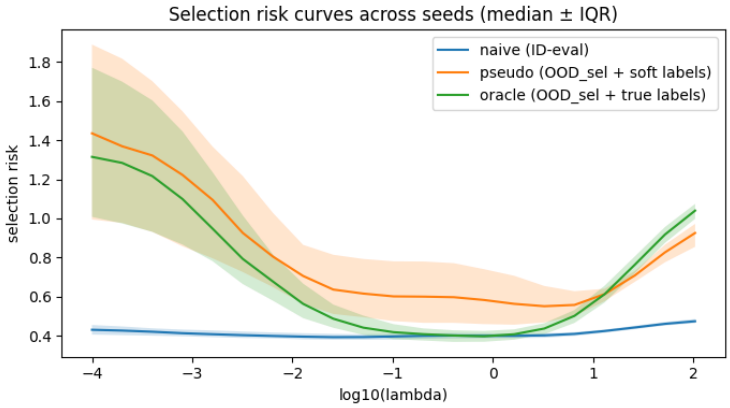


Figure 4: Selection risk curves of the three methods.

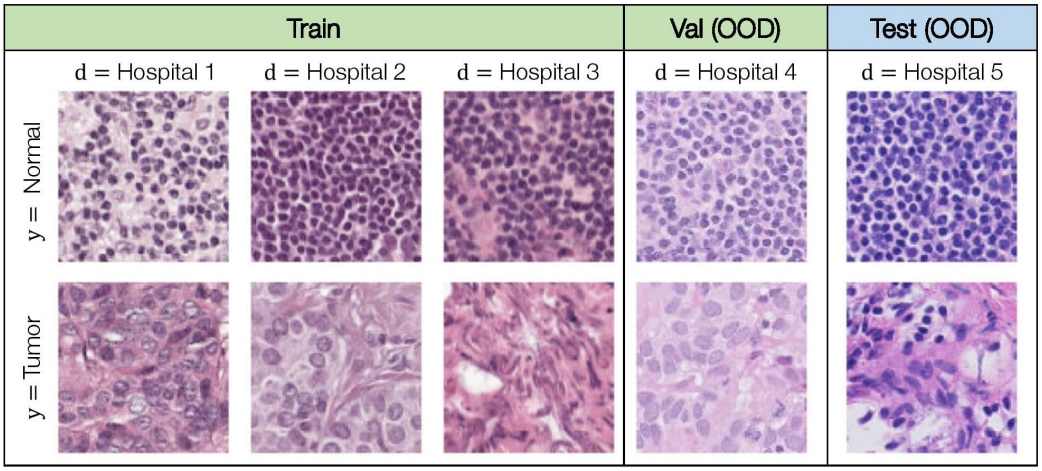


Figure 5: Camelyon17 dataset from Koh et al. (2021)

late epochs to weak regularization (full source adaptation). The first run reported below used a learning rate of 10^{-5} and $t \in \{1, \ldots, 8\}$; the two additional training seeds used a learning rate of 4×10^{-5} with a cosine schedule over 50 epochs, and were stopped after 19 and 22 epochs, respectively, beyond all epochs selected by the pseudo-labeling and oracle rules. Colab A100 GPU with high RAM runtime was used to run these experiments.

Experimental procedure. Following Algorithm 1, candidates are trained on the labeled source data $\mathcal{D}_1$, while the imputer is a separate DINOv2 instance fine-tuned to convergence on the source validation dataset $\mathcal{D}_2$, then frozen. Soft pseudo-labels are generated once on a target selection set ($\mathcal{D}_{\text{sel}}$, a random half of hospital 5) and kept fixed throughout candidate training. At each epoch t, three selection signals are computed: naive (source validation NLL on $\mathcal{D}_2$), pseudo (pseudo-target risk on $\mathcal{D}_{\text{sel}}$), and oracle (true target NLL on $\mathcal{D}_{\text{sel}}$, not available in practice). Each method selects the checkpoint minimizing its signal. Final evaluation is performed on the held-out target test set ($\mathcal{D}_{\text{test}}$, remaining half of hospital 5), which is never used during selection. No refitting is performed.

Why density ratio methods are intractable here. KMM and KLIEP estimate importance weights by solving a kernel optimization problem whose complexity scales with the number of target samples. With roughly 42k target selection set patches in a 768-dimensional DINOv2 feature space, the kernel matrix (source, target) has billions of entries, causing memory overflow even on high-RAM instances.

Subsampling the target to 2000 patches alleviates the memory issue but renders the density ratio estimate unreliable in 768 dimensions - RBF kernels with median bandwidth become flat (all pairwise distances concentrate around the same value), collapsing weights to near-uniform and providing no useful reweighting signal. This high-dimensional intractability is a known limitation of kernel density ratio methods (Sugiyama et al., 2008) and motivates our pseudo-labeling approach, which avoids density estimation entirely.

Results for the first run. Table 6 reports the three selection signals across epochs for the first run. The naive source validation loss keeps decreasing as the model fits the source data, reaching its minimum at epoch 8, by which point the model has overfit to source-specific features. Both the pseudo-target and the oracle target risks are minimized at epoch 1 and are much larger at epoch 8, so that pseudo-labeling selects the oracle checkpoint. On the target test set, the selected checkpoints reach a negative log-likelihood (NLL) of 1.942 and an accuracy of 0.531 for naive validation, against 0.655 and 0.732 for both pseudo-labeling and the oracle.

Table 6: First run: selection signals across epochs (naive: source validation NLL on $\mathcal{D}_2$; pseudo: pseudo-target risk on $\mathcal{D}_{\text{sel}}$; oracle: target NLL on $\mathcal{D}_{\text{sel}}$).

Epoch	1	2	3	4	5	6	7	8
Naive	0.111	0.071	0.056	0.057	0.041	0.045	0.045	**0.040**
Pseudo	**0.351**	0.406	0.559	0.800	0.527	0.479	0.510	1.269
Oracle	**0.659**	0.968	1.162	1.318	1.213	1.128	1.102	1.953

Stability across training seeds. We separate two independent sources of randomness: a *training seed*, controlling the head initialization, the minibatch order and the data augmentation; and a *split seed*, controlling the bisection of hospital 5 into $\mathcal{D}_{\text{sel}}$ and $\mathcal{D}_{\text{test}}$. cuDNN is deterministic and the data loader workers are seeded, so that each pair of seeds is exactly reproducible. We repeated the full fine-tuning and selection procedure for two additional training seeds. Table 7 reports the selected epochs, the target test NLL, and the regret relative to oracle selection, $\mathrm{NLL}(\text{method}) - \mathrm{NLL}(\text{oracle})$.

Table 7: Camelyon17 early stopping across three training seeds: selected epochs, target test NLL, and regret with respect to oracle selection.

	Selected epoch			Target test NLL			Regret	
Training seed	Naive	Pseudo	Oracle	Naive	Pseudo	Oracle	Naive	Pseudo
Run 1	8	1	1	1.94	0.66	0.66	1.29	0.00
Run 2	18	2	5	2.28	1.08	0.32	1.96	0.76
Run 3	21	13	2	2.05	0.71	0.31	1.74	0.40
Mean				2.09	0.82	0.43	1.66	0.39

Exact epoch-level agreement with the oracle is not robust to the initialization: pseudo-labeling selects epochs 1, 2 and 13, against oracle epochs 1, 5 and 2. Nevertheless, the practically relevant conclusion is stable. Naive source validation always selects late epochs (8, 18 and 21) at which the model has overfit the source, whereas pseudo-labeling achieves a lower target NLL than naive validation in every run, and its mean regret is more than four times smaller (0.39 versus 1.66). For Run 3, the target test accuracies are 0.695 (naive), 0.757 (pseudo-labeling) and 0.861 (oracle).

Stability across target splits. Keeping the imputer and the checkpoints of Run 3 fixed, we redrew the bisection of hospital 5 for 10 split seeds. This axis matters much less than the training seed, consistent with the size of each half ($\approx$ 42k patches): the pseudo-labeling stopping epoch lags the oracle one by exactly 11 epochs in all 10 splits, the regret, measured on the selection half, is 0.399 ± 0.004 (range $[0.394, 0.405]$), and the Spearman rank correlation between the pseudo and oracle risk curves across epochs is 0.746 ± 0.012 (mean $\pm$ standard deviation).

Scope of this experiment. As stated in Section 5.3, this experiment lies outside the scope of our theory. We do not attempt to derive a bound for this setting, as we have not carried the conditional-shift case through the estimator analysis. Note also that, with a high-capacity pretrained representation,

approximation error is unlikely to be the binding constraint: what the theory idealizes here is the shift structure rather than the model class.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 3 provides the methodology to achieve target-aware model selection with pseudo-labels under covariate shift without access to target labels, in the kernel GLM setting. Section 4 establishes target excess-risk bounds for the final estimator of the algorithm, and demonstrates its adaptivity to the unknown shift. Finally, the experiments in Section 5 show the performance gains of our methods over source-only and reweighting baselines.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion. Specific limitations are also discussed where relevant: the pure covariate shift assumption after Assumption 2.1, the requirement on the number of target covariates in Remark 4.3, the refitting used for one experiment in Section 5.2, and the scope of the Camelyon17 experiment in Section 5.3.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Our main theorem in Section 4 is accompanied by the full set of necessary assumptions (Assumptions 4.1 to 4.4), in addition to the general setup Assumption 2.1. Its proof is linked to Appendix C.3. The bulk of the appendix is dedicated to establishing this theorem, with the whole proof structure explained. In particular, a detailed proof sketch is provided in Appendix B, which also presents our two key intermediate results. A proof idea is also provided in the main body at the end Section 4.2.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Algorithm 1 presents our main algorithm and Section 5 provides the general setup of all our experiments, while Appendix H gives additional details such as hyperparameters used, and a tweaked version of our main algorithm (Algorithm 2). Full details are also provided in our GitHub repository `https://github.com/nathanweill/pseudolabel-uda-experiments`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The full code and procedure are available at `https://github.com/nathanweill/pseudolabel-uda-experiments`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: One of our core contributions is precisely the new algorithm Algorithm 1 and model selection strategy, therefore data splits and hyperparameters choice are made completely clear in Section 3. The optimization procedure and solver used are described in Appendix H.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Both our synthetic experiments in Section 5.1 and real experiments in Section 5.2 run over 100 random seeds for statistical significance, and the procedure is explained in the respective sections. Figures and tables report results with confidence interval and standard error, e.g. Table 2, Figure 1a.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute environment used to run the experiments and the approximate resources required for reproduction are reported in the GitHub repository at `https://github.com/nathanweill/pseudolabel-uda-experiments`. Most experiments are lightweight and can be reproduced on CPU; we used Google Colab Pro mainly for convenience. Experiments from Section 5.3 required using a Colab GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is methodological and uses publicly available or synthetic benchmark data. It does not involve human subjects, private or sensitive personal data, deception, or deployment in high-stakes settings. We have preserved anonymity and are not aware of any ethical concerns under the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: our work is mostly theoretical.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our work focuses on theoretical bounds and methodological improvements for kernel GLMs under covariate shift. We evaluate our approach on existing, standard benchmarks and do not release high-risk pre-trained models or newly scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets, including the Raisin and Camelyon17 datasets and baseline implementations, are appropriately cited in the main text. Details regarding their open-source licenses and terms of use are respected and documented in the README file of our GitHub repository submission (`https://github.com/nathanweill/pseudolabel-uda-experiments`).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: We do not release new datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.