Pseudo-Label Validation for Unsupervised Domain Adaptation

Nathan Weill [*1] **Kaizheng Wang** [*1 2]

Abstract

Model selection—choosing among candidate predictors, hyperparameters, or checkpoints—is a critical yet underaddressed bottleneck in unsupervised domain adaptation. Without target labels, standard validation fails, and existing unsupervised criteria rely on structural assumptions specific to feature alignment or image classification. We propose to construct a *surrogate target validation loss* by pseudo-labeling unlabeled target points with an imputation model, enabling model selection via standard holdout evaluation on the target domain. The approach is agnostic to the adaptation method, requires no density-ratio estimation, and applies to any generalized linear model loss, including logistic and square loss. Our central theoretical contribution is an oracle inequality identifying the imputer's *bias*, rather than its random error, as the dominant driver of selection quality: because a single imputer generates all pseudo-labels, its bias aggregates across validation points while its stochastic fluctuation enters only through a non-aggregating, averaged proxy. Experiments on kernel methods regularization selection and deep backbone fine-tuning confirm that surrogate validation reliably tracks target risk without any target label.

1. Introduction

The standard supervised learning framework assumes that training and test data share the same distribution (Vapnik, 1999), yet this assumption is routinely violated in practice. A canonical and pervasive failure mode is *covariate shift* (Shimodaira, 2000), where the marginal distribution of inputs differs between source and target domains while the conditional label distribution remains stable—a mismatch that can severely degrade predictive performance (Koh et al., 2021). When target labels are unavailable or costly to obtain, this motivates unsupervised domain adaptation (UDA), which must bridge the gap using only labeled source data and unlabeled target covariates. Classical UDA methods reduce the source-to-target gap through instance reweighting (Sugiyama et al., 2007; Huang et al., 2006), feature alignment (Long et al., 2015; Sun & Saenko, 2016; Pan et al., 2011), or adversarial representation learning (Ganin et al., 2016). A second wave of methods borrows from semi-supervised learning to exploit unlabeled target data (Shu et al., 2018; Prabhu et al., 2021; Berthelot et al., 2022). Recent empirical evidence suggests that the benefits of explicit adaptation methods diminish as backbone quality improves: Kalluri et al. (2024) show that with advanced pretrained representations, simple transfer baselines are highly competitive, shifting the practical problem from learning invariant features to *choosing the right predictor on top of a fixed or lightly fine-tuned backbone*.

Regardless of which adaptation strategy is adopted, every pipeline faces model-selection decisions: choosing a regularization parameter, a fine-tuning strength, or an early-stopping checkpoint. These choices are straightforward with labeled target validation, but become ill-posed without target labels. Source validation is a natural default yet leads to suboptimal selection under covariate shift. Few works propose dedicated unsupervised scorers: importance-weighted cross-validation (Sugiyama et al., 2007; You et al., 2019), prediction entropy (Morerio et al., 2018; Saito et al., 2021), or circular validation (Bruzzone & Marconcini, 2010). Recent benchmarks confirm that no existing scorer dominates across settings (Lalou et al., 2025).

We propose *pseudo-target validation*: given unlabeled target points, we construct a surrogate target validation loss by replacing missing target labels with pseudo-labels from an imputation model. The selected model is the candidate minimizing this surrogate loss. The criterion is not a new adaptation algorithm; it is a validation layer that can be placed on top of any candidate path of models. Since our criterion estimates the same target risk that a supervised target oracle would use, it is independent of cluster or neighborhood assumptions in the target representation and naturally applicable beyond classification, including general regression and generalized linear model losses.

[1]IEOR Department, Columbia University, New York, USA [2]Data Science Institute, Columbia University, New York, USA. Correspondence to: Nathan Weill <nmw2142@columbia.edu>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

Pseudo-labeling for model selection was introduced by Wang (2026) in the context of kernel ridge regression under covariate shift, whose Theorem 4.1 establishes an oracle inequality for pseudo-label model selection in the square-loss (KRR) case. Our work extends this paradigm—and that oracle inequality—to the broader and technically more demanding class of kernel GLMs, and proposes it as a general-purpose validation tool applicable beyond kernel methods to any model class. Specifically, we prove an oracle inequality showing that pseudo-target validation with GLM loss selects a near-oracle model whenever pseudo-label *bias* is small, and we derive non-asymptotic excess risk bounds for ridge-regularized kernel GLMs in the UDA setting. Finally, we validate our approach through ablations on pseudo-label regularization strength and early-stopping experiments for fine-tuning pretrained representations on the WILDS benchmark (Koh et al., 2021).

2. Unsupervised Domain Adaptation setup

2.1. Generalized linear models under covariate shift

Let $\mathcal{X}$ and $\mathcal{Y} \subseteq \mathbb{R}$ denote the input and output spaces. We observe labeled source data $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \sim \mathcal{P}^{\otimes n}$ and unlabeled target data $\mathcal{D}_0 = \{(\boldsymbol{x}_{0i}, y_{0i})\}_{i=1}^{n_0} \sim \mathcal{Q}^{\otimes n_0}$, where $\mathcal{P}$ and $\mathcal{Q}$ share the same conditional distribution but differ in their marginals over $\mathcal{X}$ (*covariate shift*). In Unsupervised Domain Adaptation (UDA), the 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 & Wedderburn, 1972) in a generic function class $\mathcal{F}$. Specifically, conditioned on $\boldsymbol{x}$, the response y is drawn from an exponential family distribution with log-likelihood (up to constants):

$$\ell(y; f^*(\boldsymbol{x})) = y f^*(\boldsymbol{x}) - a(f^*(\boldsymbol{x})), \tag{2.1}$$

where $a : \mathbb{R} \to \mathbb{R}$ is a strictly convex log-partition function, and $f^* \in \mathcal{F}$ is the true underlying signal (well-specified model). The conditional mean of the response is given by the canonical link:

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

Three canonical instances:

- **Least-squares regression** ($a(t) = t^2/2$): recovers the squared loss $(f(\boldsymbol{x}) - y)^2/2$, with $\mathbb{E}[y \mid \boldsymbol{x}] = f(\boldsymbol{x})$.
- **Binary classification** ($a(t) = \log(1 + e^t)$): recovers the logistic loss, with $\mathbb{E}[y \mid \boldsymbol{x}] = \sigma(f(\boldsymbol{x})) \in (0, 1)$ [1].
- **Count regression** ($a(t) = e^t$): recovers the Poisson deviance loss, with $\mathbb{E}[y \mid \boldsymbol{x}] = e^{f(\boldsymbol{x})}$.

[1] σ is the sigmoid function.

2.2. The Model Selection Problem

We are given a finite set of candidate predictors $\{f_1, \ldots, f_m\} \subseteq \mathcal{F}$—for instance, checkpoints along a training trajectory, models trained with different regularization strengths, or outputs of different adaptation methods. The goal is to identify the candidate with the best target performance using only the unlabeled target data $\mathcal{D}_0$.

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

$$\begin{aligned} R(f) &= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{Q}}[a(f(\boldsymbol{x})) - y f(\boldsymbol{x})] \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mathcal{Q}}[a(f(\boldsymbol{x})) - f(\boldsymbol{x}) a'(f^*(\boldsymbol{x}))]. \end{aligned} \tag{2.3}$$

3. Surrogate Target Validation Loss

In standard hold-out validation, with access to target labels $\{y_{0i}\}_{i=1}^{n_0}$, we select the parameter λ minimizing the empirical target risk on an independent validation set:

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

This is already a noisy version of the "oracle" empirical risk

$$R_{in}(f) = \frac{1}{n_0} \sum_{i=1}^{n_0} \left(a(f(\boldsymbol{x}_{0i})) - a'(f^*(\boldsymbol{x}_{0i})) f(\boldsymbol{x}_{0i}) \right), \tag{3.2}$$

yet validation risk is a reliable guide to generalization. Due to missing labels, we require a strategy to estimate the target performance using only the unlabeled covariates.

We propose to construct a computable proxy for the empirical risk (3.2). 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 (3.1).

We define the *surrogate target validation loss* of a candidate f as

$$\widehat{R}_{in}(f) = \frac{1}{n_0} \sum_{i=1}^{n_0} \left(a(f(\boldsymbol{x}_{0i})) - \tilde{y}_{0i} f(\boldsymbol{x}_{0i}) \right). \tag{3.3}$$

Model selection then proceeds by $\hat{f} = \arg\min_{j\in[m]} \widehat{R}_{in}(f_j)$, or, in the case of a training trajectory, by monitoring $\widehat{R}_{in}$ over iterations and applying early stopping. Importantly, if target data was also used during adaptation, the validation set $\mathcal{D}_0$ must be held out from that process to avoid overfitting.

Pseudo-labels are mean predictions. Pseudo-labels are defined as the *predicted conditional mean* of y under the imputation model. In all cases $\tilde{y}_{0i} = a'(\tilde{f}(\boldsymbol{x}_{0i}))$ is a *deterministic* quantity on the response scale—the imputer's estimate of $\mathbb{E}[y \mid \boldsymbol{x}_{0i}]$—and not a resampled or re-imputed label; the only randomness our analysis attributes to the pseudo-labels enters through the fitted imputer $\tilde{f}$. While in regression, $\tilde{y}_{0i} = \tilde{f}(\boldsymbol{x}_{0i})$ is the direct real-valued prediction, in classification, our method utilizes "soft" pseudo-labels $\tilde{y}_{0i} = \sigma(\tilde{f}(\boldsymbol{x}_{0i})) \in (0,1)$—representing the estimated conditional probability—rather than "hard" labels $\mathbf{1}(\tilde{f}(\boldsymbol{x}_{0i}) \geq 0)$, which would destroy the calibration information necessary for minimizing the log-likelihood.

Intuitively, as long as the imputer $\tilde{f}$ captures the conditional mean $a'(f^*(\boldsymbol{x}))$ reasonably well on target points, the ranking induced by $\widehat{R}_{in}$ will track the true ranking induced by R_{in}. Our theory in Section 4 formalises this intuition and shows that the relevant bias–variance trade-off for the imputer is *asymmetric*: its bias aggregates over target points while its random error does not, so the optimum lies on the undersmoothed side—the imputer should carry less regularization than a prediction-optimal fit would use. The penalty cannot be driven to zero, however, since the imputer must stay consistent; the optimum therefore sits at the smallest penalty for which it remains well-behaved. This prescription is validated empirically through an ablation over the imputer's regularization strength in kernel GLMs in Appendix I.3.

Given labeled source data $\mathcal{D}$ and unlabeled target data $\mathcal{D}_0$, the general UDA model selection methodology will thus be:

1. Split $\mathcal{D}$ into disjoint folds $\mathcal{D}_1$ and $\mathcal{D}_2$. Use $\mathcal{D}_1$ to train a family of candidate models $\{f_1, \ldots, f_m\}$, and $\mathcal{D}_2$ to train an undersmoothed imputation model $\tilde{f}$.
2. Compute pseudo-labels for the target covariates on $\mathcal{D}_0$ via $a'(\tilde{f}(\boldsymbol{x}_{0i}))$. Select the candidate minimizing the surrogate target validation loss (3.3): $\hat{f} = \arg\min_{j\in[m]} \widehat{R}_{in}(f_j)$.

4. Generic model selection analysis with pseudo-labels under the GLM loss

We need to understand how the imperfection of pseudo-labeler impacts the model selection. To isolate the effect of pseudo-label errors from the finite-sample fluctuations of the target covariates, we treat in this section 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 decompose the pseudo-labels as

$$\tilde{f}(\boldsymbol{x}_{0i}) := f^*(\boldsymbol{x}_{0i}) + B_i + \tilde{\varepsilon}_i, \tag{4.1}$$

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

In this section, we will equivalently work with the excess risk versions of Eq. (3.2) and Eq. (3.3). The oracle in-sample excess risk is given by

$$\mathcal{R}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n} \Big[a(f(\boldsymbol{x}_{0i})) - a(f^*(\boldsymbol{x}_{0i})) - a'(f^*(\boldsymbol{x}_{0i})(f(\boldsymbol{x}_{0i}) - f^*(\boldsymbol{x}_{0i}))\Big], \tag{4.2}$$

and the surrogate is

$$\widehat{\mathcal{R}}_{in}(f) = \frac{1}{n}\sum_{i=1}^{n} \Big[a(f(\boldsymbol{x}_{0i})) - a(\tilde{f}(\boldsymbol{x}_{0i})) - a'(\tilde{f}(\boldsymbol{x}_{0i})(f(\boldsymbol{x}_{0i}) - \tilde{f}(\boldsymbol{x}_{0i}))\Big]. \tag{4.3}$$

The pseudo-labeling strategy is thus to choose $\hat{j} = \operatorname{argmin}_{j\in[m]} \hat{\mathcal{R}}_{in}(f_j)$ and we denote the selected model by $\hat{f} = f_{\hat{j}}$.

Assumption 4.1. *We impose some regularity conditions on a and f^*, assume that the imputer has low pointwise error on the target, and that the imputation error is sub-exponential.*

1. *The image of the covariates under the true model is bounded almost surely, i.e. there exists a constant $C > 0$ such that $|f^*(\boldsymbol{x}_{0i})| \leq C$, $\forall i$.*
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 $\forall u, v \in [-2C, 2C]$, $K \geq a''(u) \geq k$ and $|a''(u) - a''(v)| \leq L|u - v|$.*
3. *With probability at least $1 - \delta/2$, $\max_{i\in[n]} |\tilde{f}(\boldsymbol{x}_{0i}) - f^*(\boldsymbol{x}_{0i})| \leq \eta$, for a constant $\eta < \min\{k/L, C, \sqrt{kK}/L\}$.*
4. *The imputation random error $\tilde{\boldsymbol{\varepsilon}} = (\varepsilon_1, ..., \varepsilon_n)^\top$ 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 true 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 $j^* = \operatorname{argmin}_{j\in[m]} \mathcal{R}_{in}(f_j)$. To lighten the notations, denote in what follows $\mathcal{R}_{in}(f_j) = \mathcal{R}_j, \hat{\mathcal{R}}_{in}(f_j) = \hat{\mathcal{R}}_j, \forall j$.

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

Theorem 4.1. *Let $\delta \in (0,1)$. Then there exists a universal constant $C_1 > 0$ such that under Assumption 4.1, the following holds with probability at least $1-\delta$:*

$$\mathcal{R}_{\hat{j}} \leq \inf_{\epsilon>0} \left\{ (1+\epsilon)\mathcal{R}_{j^*} + \frac{C_1 K}{nk}\left(1+\frac{1}{\epsilon}\right)\xi(\tilde{\sigma}, \boldsymbol{B}, \eta, \delta) \right\} \tag{4.4}$$

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

The proof is deferred to Appendix G. The additive overhead in the oracle inequality (4.4) corresponds to the error of the pseudo-labels. It comprises a bias term proportional to $\|\boldsymbol{B}\|^2/n$, a variance term proportional to $\tilde{\sigma}^2 \log^2(m)/n$, and a small perturbation term proportional to $L^2\eta^4$. 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. We stress that this is *not* the classical averaging of independent label noise—all errors $\tilde{\varepsilon}_i$ come from the single imputer $\tilde{f}$ and are correlated across target points—so the asymmetry is structural rather than a noise-cancellation effect.

This asymmetry is clearest when the overhead is placed beside the imputer's own excess risk. Decomposing both informally at the design points into the same bias vector $\boldsymbol{B}$ and noise vector $\tilde{\boldsymbol{\varepsilon}}$,

$$\text{imputer's excess risk} \;\lesssim\; \tfrac{1}{n}\big(\,\|\boldsymbol{B}\|_2^2 + \|\tilde{\boldsymbol{\varepsilon}}\|_2^2\big),$$

$$\text{selection overhead} \;\asymp\; \tfrac{1}{n}\big(\,\|\boldsymbol{B}\|_2^2 + \tilde{\sigma}^2 \log^2 m\big).$$

the bias term $\|\boldsymbol{B}\|_2^2$ is *identical*; only the noise differs. The imputer's excess risk pays the full energy $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2$, which can be as large as $n\tilde{\sigma}^2$ (one contribution per coordinate), whereas the selection overhead pays only the scalar proxy $\tilde{\sigma}^2$—smaller by a factor of n. Selection thus inherits the imputer's bias but discounts its variance by n.

Consequently, as long as $\log^2(m)/n$ is small, even a moderate pointwise bias dominates the variance contribution, making bias the critical quantity to control. The two uses of the imputer therefore call for different amounts of regularization: whereas minimizing its own excess risk dictates a standard bias–variance balance, using it for selection tolerates—and indeed benefits from—a more aggressively *undersmoothed* fit with smaller bias, so that the pseudo-labels select a more faithful model. We show in Appendix B that, once the imputer is instantiated for kernel GLMs, the optimized trade-off leaves a pseudo-label overhead that is lower-order relative to the candidate risk.

Table 1. Risk mean estimate across seeds with confidence interval (CI) and standard error (SE).

Method	Mean	95% CI	SE
Naive	0.502	[0.463, 0.541]	0.019
KLIEP + IW-CV	0.478	[0.438, 0.518]	0.020
KMM + IW-CV	0.469	[0.442, 0.496]	0.014
Pseudo-labeling	0.428	[0.412, 0.444]	0.008
Oracle	0.373	[0.361, 0.385]	0.006

5. Examples of application

We instantiate the surrogate target validation framework in two settings that together illustrate its scope: regularization selection for kernel GLMs, and early stopping for deep network fine-tuning on a real distribution shift benchmark, where the method is used as a practical tool. We evaluate the selected models on a held-out target test dataset with known labels $\mathcal{D}_{\text{test}}$. The code for implementing our experiments is available at https://anonymous.4open.science/r/pseudo-label-validation-uda-2601.

5.1. Regularization Selection for Kernel GLMs

Setup. Let $\mathcal{F}$ be a reproducing kernel Hilbert space (RKHS) associated with a kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The candidate models are a family of kernel GLM estimators indexed by a positive ridge regularization strength $\{\hat{f}_\lambda\}_{\lambda\in\Lambda}$ trained on $\mathcal{D}_1$ where

$$\hat{f}_\lambda = \operatorname*{argmin}_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^n \left(a(f(\boldsymbol{x}_i)) - y_i f(\boldsymbol{x}_i)\right) + \frac{\lambda}{2}\|f\|_{\mathcal{F}}^2 \right\}, \tag{5.1}$$

and the imputer is itself a kernel GLM trained on $\mathcal{D}_2$ with a fixed, small regularization $\tilde{\lambda}$. The full procedure is described in Appendix B, along with non-asymptotic excess risk bounds for the final estimator.

Experiments. Experiments are conducted on the Raisin dataset (Çınar et al., 2020), that contains 900 instances of two raisin varieties with 7 numerical features. The task is binary classification and we introduce covariate shift via a usual sub-sampling technique on the first covariate (Quiñonero-Candela et al., 2008). Table 1 reports the validation results under our default protocol—repeated 2-fold cross-validation, whose per-fold split allocates half the source data to the candidates and half to the imputer, matching the $n_1 = n_2 = n/2$ split analyzed in Algorithm 1—comparing pseudo-label surrogate selection with naive source-only validation, oracle selection with labeled target data, and importance-weighted baselines. Full details and an ablation on the split ratio can be found in Appendix I.4.

5.2. Early Stopping for Fine-Tuned DINOv2 on Camelyon17

We consider early stopping determination for fine-tuning on the Camelyon17 (Bandi et al., 2018) benchmark from WILDS (Koh et al., 2021). It is a large-scale medical imaging dataset, which frames tumor detection in 96x96 histopathological patches as a binary classification task. A natural domain shift occurs across hospitals because of changes in patient population and image acquisition. A pre-trained DINOv2 ViT-B/14 (Oquab et al., 2023) backbone is fine-tuned end-to-end on $\mathcal{D}_1$ ($\approx$ 300k samples from hospitals 0–3), with the epoch number serving as the regularization parameter, while the imputer is trained on $\mathcal{D}_2$ (≈ 34k samples) and generates pseudo-labels on $\mathcal{D}_0$ (≈ 42k samples from hospital 4); the epoch minimizing pseudo-target risk is selected and the associated checkpoint is evaluated on $\mathcal{D}_{\text{test}}$ (≈ 42k samples). Naive source validation selects epoch 8, by which point the model has catastrophically overfit to source-specific features, achieving near-random target accuracy (53%) and a log-loss of 1.94. Conversely, our pseudo-labeling method correctly identifies epoch 1 as optimal, perfectly matching the oracle and yielding a 73% target accuracy and a log-loss of 0.66. Full experimental details, hyperparameters, and additional diagnostics are provided in Appendix I.5.

6. Conclusion

We have proposed pseudo-target validation, a simple and principled approach to model selection in unsupervised domain adaptation: replace unavailable target labels with pseudo-labels from an imputation model and evaluate candidates via a surrogate target validation loss computed entirely from unlabeled data. Our oracle inequality identifies pseudo-label bias as the main driver of selection error—making the imputer's bias–variance trade-off asymmetric, with the optimum on the undersmoothed side—and experiments on kernel GLMs and deep backbone fine-tuning on WILDS confirm that surrogate validation reliably tracks true target risk without any target label. Future directions include applying pseudo-target validation to broader hyperparameter selection problems—weight decay, LoRA rank, number of frozen layers, or distillation temperature—where principled surrogates are currently lacking.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

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

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., 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.

Berthelot, D., Roelofs, R., Sohn, K., Carlini, N., and Kurakin, A. AdaMatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations (ICLR)*, 2022.

Bruzzone, L. and Marconcini, M. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):770–787, 2010. doi: 10.1109/TPAMI.2009.57.

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

Charlier, B., Feydy, J., Glaunès, J., Collin, F.-D., and Durif, G. Kernel operations on the GPU, with autodiff, without memory overflows. *Journal of Machine Learning Research*, 22(74):1–6, 2021.

Çınar, İ., Köklü, M., and Taşdemir, Ş. Raisin. UCI Machine Learning Repository, 2020.

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

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Hestenes, M. R. and Stiefel, E. 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.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., and Smola, A. Correcting sample selection bias by unlabeled data. In Schölkopf, B., Platt, J., and Hoffman, T. (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Kalluri, T., Ravichandran, S., and Chandraker, M. UDA-Bench: Revisiting common assumptions in unsupervised domain adaptation using a standardized framework. In *Computer Vision – ECCV 2024*, volume 15143 of *Lecture Notes in Computer Science*, pp. 196–213. Springer, 2024. doi: 10.1007/978-3-031-73013-9_12.

Kim, J.-H. 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.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B. A., Haque, I. S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., and Liang, P. 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*, pp. 5637–5664. PMLR, 2021.

Lalou, Y., Gnassounou, T., Collas, A., de Mathelin, A., Kachaiev, O., Odonnat, A., Gramfort, A., Moreau, T., and Flamary, R. Skada-bench: Benchmarking unsupervised domain adaptation methods with realistic validation on diverse modalities, 2025.

Long, M., Cao, Y., Wang, J., and Jordan, M. I. Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 97–105. PMLR, 2015.

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

Marteau-Ferey, U., Ostrovskii, D., Bach, F., and Rudi, A. 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*, pp. 2294–2340. PMLR, 2019. doi: 10.48550/arxiv.1902.03046.

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

Morerio, P., Cavazza, J., and Murino, V. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *International Conference on Learning Representations*, 2018.

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

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision, 2023.

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. doi: 10.1109/TNN.2010.2091281.

Prabhu, V., Khare, S., Kartik, D., and Hoffman, J. SENTRY: Selective entropy optimization via committee consistency for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8558–8567, 2021.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (eds.). *Dataset Shift in Machine Learning*. The MIT Press, 2008. ISBN 9780262255103. doi: 10.7551/mitpress/9780262170055.001.0001.

Saito, K., Kim, D., Teterwak, P., Sclaroff, S., Darrell, T., and Saenko, K. Tune it the right way: Unsupervised validation of domain adaptation via soft neighborhood density. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9184–9193, 2021.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

Shimodaira, H. 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.

Shu, R., Bui, H. H., Narui, H., and Ermon, S. A DIRT-T approach to unsupervised domain adaptation. In *International Conference on Learning Representations*, 2018.

Sugiyama, M., Krauledat, M., and Müller, K.-R. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8:985–1005, 2007.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. 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.

Sun, B. and Saenko, K. Deep CORAL: Correlation alignment for deep domain adaptation. In *Computer Vision – ECCV 2016 Workshops*, volume 9915 of *Lecture Notes in Computer Science*, pp. 443–450. Springer, 2016. doi: 10.1007/978-3-319-49409-8_35.

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

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

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

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

You, K., Long, M., Cao, Z., Wang, J., and Jordan, M. I. Towards accurate model selection in deep unsupervised domain adaptation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7124–7133. PMLR, 2019.

Appendices

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

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

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

We define the *excess risk* as:

$$\begin{aligned}
\mathcal{R}(f) &= R(f) - R(f^*) \\
&= \mathbb{E}_{\boldsymbol{x} \sim \mathcal{Q}} \Big[a(f(\boldsymbol{x})) - a(f^*(\boldsymbol{x})) \\
&\quad - a'(f^*(\boldsymbol{x}))(f(\boldsymbol{x}) - f^*(\boldsymbol{x})) \Big].
\end{aligned} \tag{A.1}$$

Note that, on the domain where Assumption B.5 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:

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

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

Likewise, for the finite-sample version, we get

$$\frac{k}{2n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i)-f^*(\boldsymbol{x}_i)\right)^2 \leq \mathcal{R}_{in}(f) \leq \frac{K}{2n}\sum_{i=1}^{n}\left(f(\boldsymbol{x}_i)-f^*(\boldsymbol{x}_i)\right)^2. \tag{A.3}$$

Remark A.2 (Excess risk bounds)**.** Based on Eqs. (A.2) and (A.3) and the notations introduced, we have for any $\lambda > 0$:

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

and

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

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. RKHS, Kernel examples and ridge estimation

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:

$$\mathcal{F} = \{f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R} \mid f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \langle\Phi(\boldsymbol{x}),\boldsymbol{\theta}\rangle_{\mathbb{H}},\ \boldsymbol{\theta} \in \mathbb{H}\}. \tag{A.4}$$

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

Below are some frequently used kernels and associated RKHS's (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

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

By the Representer Theorem (Schölkopf et al., 2001), the solution of the kernel ridge GLM problem presented in Eq. (5.1) 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:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^n} \left\{ \frac{1}{n} \left(\mathbf{1}^\top \boldsymbol{a}(\boldsymbol{\alpha}) - \boldsymbol{y}^\top \boldsymbol{K}\boldsymbol{\alpha} \right) + \frac{\lambda}{2} \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha} \right\}, \tag{A.5}$$

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

B. Kernel GLM example: assumptions and theoretical guarantees

This section develops theoretical guarantees for Algorithm 1 below. 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.

Algorithm 1 Pseudo-Labeling for Kernel GLM

1: **Input**: Source data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, unlabeled target covariates $\{\boldsymbol{x}_{0i}\}_{i=1}^{n_0}$, 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 minimizing the surrogate target validation loss (3.3) evaluated at each $\hat{f}_\lambda$:

$$\hat{\lambda} \in \operatorname*{argmin}_{\lambda\in\Lambda} \widehat{R}_{in}(\hat{f}_\lambda) = \operatorname*{argmin}_{\lambda\in\Lambda} \frac{1}{n_0} \sum_{i=1}^{n_0} \left(a(\hat{f}_\lambda(\boldsymbol{x}_{0i})) - \tilde{y}_{0i} \hat{f}_\lambda(\boldsymbol{x}_{0i}) \right).$$

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

Remark B.1 (On the data split)**.** For simplicity our analysis fixes a balanced split $n_1 = n_2 = n/2$ (Algorithm 1). This ratio is not optimized: the candidate excess risk depends on n_1 (through n_{eff}) while the pseudo-label overhead depends on n_2, so an asymmetric split could in principle trade these two contributions off. Sample splitting also forfeits a constant factor relative to reusing all data for both roles; this loss can be recovered by *cross-fitting*—swapping the candidate and imputer halves and averaging the resulting risk estimates—which is exactly the repeated K-fold procedure we adopt in practice (Algorithm 2). A formal optimization of the split ratio and a cross-fitted analysis are left to future work.

B.1. Assumptions

Assumption B.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* $\operatorname{supp}(\mathcal{Q}) \subseteq \operatorname{supp}(\mathcal{P})$. *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.*

Our analysis of kernel ridge GLM under covariate shift (Assumption B.1) necessitates only mild regularity conditions, that one can verify are satisfied in most common scenarios.

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 B.2 (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 admits a Mercer expansion under each marginal,*

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^{\infty} \mu_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{z}) \quad \text{in } L^2(\mathcal{X}; \mathcal{P}), \qquad K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^{\infty} \mu_{0j} \phi_{0j}(\boldsymbol{x}) \phi_{0j}(\boldsymbol{z}) \quad \text{in } L^2(\mathcal{X}; \mathcal{Q}),$$

  i.e. two distinct eigen-decompositions, each adapted to its own distribution, that represent the same kernel, where $(\mu_j)_{j=1}^{\infty}$ and $(\mu_{0j})_{j=1}^{\infty}$ are sequences of positive eigenvalues [2], $(\phi_j)_{j=1}^{\infty}$ 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.

- $\exists M_1 < \infty, \forall j, \|\phi_j\|_\infty \leq M_1, \|\phi_{0j}\|_\infty \leq M_1$.

- $\exists A, A' > 0$, $\exists \alpha \geq 1$, $\forall j \geq 1, \mu_j \leq A j^{-2}$, $\mu_{0j} \leq A' j^{-2\alpha}$.

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 $\mathcal{X} = [0, 1]$ (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, whose eigenfunctions are sinusoidal.

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 B.3 (Sub-exponential noise)**.** *Conditioned on $\boldsymbol{X}$, the noise vector $\boldsymbol{\varepsilon}$ satisfies $\|\boldsymbol{\varepsilon}\|_{\psi_1} \leq \sigma < \infty$.*

Assumption B.4 (Bounded signal)**.** $\|f^*\|_{\mathcal{F}} \leq R < \infty$.

Under Assumption B.2, the feature map induced by the kernel K can be written as

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

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 B.1, $\forall \boldsymbol{x} \in \operatorname{supp}(\mathcal{P}), |f^*(\boldsymbol{x})| \leq C := RM_2$.

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

[2] 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.

Table 2. 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}

Assumption B.5 (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 $k_\rho := \inf_{|z|\le(\rho+1)C} a''(z) > 0$. Define $k = k_1$. Moreover, a'' is assumed to be upper bounded and Lipschitz on compact regions: let $K_\rho := \sup_{|z|\le(\rho+1)C} a''(z) < \infty$, $L_\rho := \sup_{|z|\le(\rho+1)C} a'''(z) < \infty$, and define $K = K_1$, $L = L_1$. Then, $\forall |z| \le 2C, k \le a''(z) \le K$ and $\forall |z_1|, |z_2| \le 2C, |a''(z_1) - a''(z_2)| \le L\,|z_1 - z_2|$.*

Table 2 provides simple bounds for the corresponding GLM log-partition constants in the three canonical examples.

B.2. Main results

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

Theorem B.1 (Excess risk)**.** *Let Assumptions B.1 to B.5 hold. Define for $\mu^2 = M_1^6\sigma^2$*

$$n_{\mathit{eff}} = \sup\left\{ t \le n \;\middle|\; t\boldsymbol{\Sigma}_0 \preceq n\boldsymbol{\Sigma} + \frac{\mu^2}{k}\boldsymbol{I} \right\}. \tag{B.1}$$

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

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

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 $1-\delta$:

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

The proof is deferred to Appendix D.3.

Remark B.2 (Interpretation of Effective Sample Size)**.** The quantity n_{eff} defined in Eq. (B.1) is a data-dependent measure of how informative the source data are for the target task. It is primarily governed by the covariate shift—the extent to which the source covariance $\boldsymbol{\Sigma}$ "covers" the target covariance $\boldsymbol{\Sigma}_0$—but is not a pure covariate-shift functional: through $\mu^2 = M_1^6\sigma^2$ it also carries a noise-scale term μ^2/k, the regularization floor imposed by the imputer's stochastic error, which prevents directions with negligible source variance from inflating n_{eff} arbitrarily. The covariate-shift part of 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.

- If the source dominates the target (i.e., $\boldsymbol{\Sigma}_0 \preceq B\boldsymbol{\Sigma}$), we have $n_{\text{eff}} \ge n/B$, recovering standard fast rates.
- If the target $\boldsymbol{\Sigma}_0$ has high energy in eigen-directions where the source $\boldsymbol{\Sigma}$ 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 (B.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 (B.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 B.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 $1 - n^{-1}$:*

$$R(\hat{f}) - R(f^*) \lesssim n_{\mathit{eff}}^{-\frac{2\alpha}{2\alpha+1}} + n_0^{-1}. \tag{B.3}$$

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

Remark B.3 (Optimality and adaptivity)**.** The bound established in Corollary B.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.

B.3. Proof structure

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

C. Proof sketch and key intermediate results

Our main result hinges on two ingredients. Section 4 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 C.1 then shows how, in the 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 C.1 as a key intermediate result. Appendix C.2 explains how these two ingredients are combined to yield Theorem B.1.

C.1. Analysis of the KRGLM estimator

We now show how the oracle inequality of Theorem 4.1 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. (5.1) 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. (5.1) becomes

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

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

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

where the data-dependent Hessian is

$$\boldsymbol{H} = \frac{1}{n}\sum_{i=1}^n D_{ii}\, \boldsymbol{x}_i \boldsymbol{x}_i^\top + \lambda \boldsymbol{I}, \qquad D_{ii} = \int_0^1 a''(\boldsymbol{x}_i^\top(\boldsymbol{\theta}^* + t(\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^*)))\, dt. \tag{C.3}$$

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 (C.2) accordingly gives the decomposition:

$$\hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* = -\Bigg[\underbrace{\lambda \boldsymbol{H}^{-1}\boldsymbol{\theta}^*}_{\text{bias}} + \underbrace{\frac{1}{n}\bar{\boldsymbol{H}}^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}}_{\text{variance}} + \underbrace{\frac{1}{n}(\boldsymbol{H}^{-1} - \bar{\boldsymbol{H}}^{-1})\boldsymbol{X}^\top \boldsymbol{\varepsilon}}_{\text{perturbation}}\Bigg]. \tag{C.4}$$

An equivalent decomposition holds in the general setting via the isometry between the RKHS $\mathcal{F}$ and the feature space $\mathbb{H}$ Eq. (A.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 4.1 applied to $\mathcal{D}_0$, we project decomposition Eq. (C.4) 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 F.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 B.5 — 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 C.1 (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. (5.1). Let $\beta \in (0, 8)$, and choose $\delta \in (0, 1/e]$. Then, under Assumptions B.1 to B.5, there exist constants $c, c', c_1, c_2 > 0$, such that for any given $\boldsymbol{x}_0 \in \operatorname{supp}(\mathcal{P})$, 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 $1-\delta$:

$$\left|\hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0)\right| \le cM_1\left(\sigma \log(\frac{n}{\delta})\frac{\lambda^{-1/4}}{\sqrt{n}} + \lambda^{1/4}\left\|f^*\right\|_{\mathcal{F}}\right). \tag{C.5}$$

When setting λ to its lower bound, the result becomes

$$\left|\hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0)\right| \le c'\sqrt{M_1^3\sigma}\, n^{-1/4}\left(\log(\frac{n}{\delta})\right)^3. \tag{C.6}$$

The proof is deferred to Appendix F.

Bounding the oracle inequality terms via the pointwise rate. We can now carry out the program described above. Projecting (C.4) and applying Theorem C.1 to control the perturbation yields:

1. **Candidate excess risk bounds.** Contracting (C.4) 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:

$$\min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} \lesssim \lambda R^2 \left\|\boldsymbol{S}_\lambda\right\| + \frac{\sigma^2}{n_1}\operatorname{Tr}(\boldsymbol{S}_\lambda), \tag{C.7}$$

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 D.3. Using the effective sample size to relate $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}_0$, one can easily verify that $\operatorname{Tr}(\boldsymbol{S}_\lambda) \le \frac{2n}{n_{\text{eff}}}\operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \lambda\frac{n}{n_{\text{eff}}}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0)$ and $\|\boldsymbol{S}_\lambda\| \le \frac{n}{kn_{\text{eff}}}$, as shown in Appendix H.3. Thus taking $n_1 = n/2$, it reduces to

$$\min_{\lambda\in\Lambda}\{R(\hat{f}_\lambda) - R(f^*)\} \lesssim \frac{n}{n_{\text{eff}}}\lambda R^2 + \frac{\sigma^2}{n_{\text{eff}}}\operatorname{Tr}((k\boldsymbol{\Sigma}_0 + \frac{n}{n_{\text{eff}}}\lambda\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0), \tag{C.8}$$

as proved in Appendix D.3.

2. **Imputation error decomposition.** Following the insight of Section 4, 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 C.1. Applying (C.4) 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 $\xi(\tilde{\sigma}, \boldsymbol{B}, \eta, \delta)$ in Theorem 4.1, and allows to bound them in Lemma D.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:

$$\tilde{\sigma}^2 \lesssim \frac{n_0}{n_2} \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\|, \quad \|\boldsymbol{B}\|^2 \lesssim \frac{n_0}{n_2} \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\|, \tag{C.9}$$

where $\hat{\boldsymbol{S}}_{2,\tilde{\lambda}} = (k\hat{\boldsymbol{\Sigma}}_2 + \tilde{\lambda}\boldsymbol{I})^{-1/2}\hat{\boldsymbol{\Sigma}}_0(k\hat{\boldsymbol{\Sigma}}_2 + \tilde{\lambda}\boldsymbol{I})^{-1/2}$. This equal order does not contradict the bias-driven narrative of Theorem 4.1. The asymmetry there acts *upstream*, in the choice of $\tilde{\lambda}$: since the bias aggregates over design points while the variance proxy does not, the overhead is minimized by undersmoothing, i.e. by pushing $\tilde{\lambda}$ to the lower end of the range admissible under Theorem C.1 (here $\tilde{\lambda} \asymp n^{-1}$). At that $\tilde{\lambda}$ the two terms coincide in order and together contribute only the negligible $O(1/n_{\text{eff}} + 1/n_0)$ overhead, so the kernel-GLM rate is governed by the candidate complexity rather than by a bias-versus-variance competition in the imputer. 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 C.1 yields

$$n_0\eta^4 \lesssim \frac{n_0}{n_2}. \tag{C.10}$$

Pointwise rate. We conclude this section by commenting on the rate obtained in Theorem C.1. 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 & 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 D.5 how to recover the optimal prediction rate $\left\|\hat{f}_\lambda - f^*\right\|_{L^2} = O(n^{-1/3})$ 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.

C.2. Putting things together

We now combine the two previous ingredients. First, we can substitute the bounds on the imputation error decomposition from Eqs. (C.9) and (C.10) with $n_2 = n/2$ in the oracle inequality of Theorem 4.1. 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 D.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 D.2. To obtain Theorem B.1, it then suffices to substitute the candidate excess risk bounds from Eq. (C.8) into the first term. As seen in Corollary B.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.

D. Proof of Theorem B.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.

D.1. Fixed-design analysis

Lemma D.1 (Fixed-design KRGLM)**.** *Let Assumptions B.1 to B.5 hold except that all covariates are deterministic and the data split is fixed. Choose any λ satisfying Theorem C.1 conditions:*

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

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 $1 - \delta$:

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

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

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

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 $\boldsymbol{X}_0(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{B}_1 + \boldsymbol{B}_2 + \tilde{\varepsilon}$, respectively bias, perturbation and variance, satisfying simultaneously, with probability at least $1 - \delta/2$, for all $\delta \in (0, 1/e]$:

- $\|\boldsymbol{B}_1\|_2^2 \leq n_0 \tilde{\lambda} R^2 \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\| \leq C M_1^2 \sigma^2 R^2 \frac{n_0}{n_2} \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\| \log(\frac{n_0}{\delta})(\log n_2)^6$,
- $\|\boldsymbol{B}_2\|_2^2 \leq C M_1^4 \sigma^2 \frac{L^2 n_0}{k^2 n_2} \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\| \log^7(n_2/\delta)$,
- $\|\tilde{\varepsilon}\|_{\psi_1}^2 \leq \frac{n_0}{n_2} \frac{\sigma^2}{k} \left\| \hat{\boldsymbol{S}}_{2,\tilde{\lambda}} \right\|$ *(this is deterministic), and* $\|\tilde{\varepsilon}\|_2^2 \leq C \frac{n_0}{n_2} \frac{\sigma^2}{k} \operatorname{Tr}(\hat{\boldsymbol{S}}_{2,\tilde{\lambda}}) \log^3(n_2/\delta)$.

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

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

Lemma D.2 (Concentration of second-moment operators)**.** *For any λ satisfying Theorem C.1 conditions when $n_1 = n_2 = n/2$:*

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

, and $\lambda_0 = \frac{16 M_2^2 \log(28 n_0/\delta)}{n_0}$, 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

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

D.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 E.3.

Lemma D.3 (Excess risk of the best candidate)**.** *Let Assumptions B.1 to B.5 hold. Choose any $\delta \in (0, 1/e]$, and define for any $\lambda > 0$*

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

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 4.1 combined with Lemma D.1 yields the following result for the in-sample risk, and the proof is in Appendix E.4.

Lemma D.4 (Oracle inequality for in-sample risk)**.** *Let Assumptions B.1 to B.5 hold, except that all covariates are deterministic and the data split is fixed. Choose any $\delta \in (0, 1/e]$, and let $\tilde{\lambda} = \frac{M_1^2 \sigma^2 \log(16 n_0/\delta)(\log n_2)^{2+\beta/2}}{n_2}$, then define*

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

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 E.5.

Lemma D.5 (Bridging in-sample and out-of-sample risks)**.** *Let Assumptions B.1 to B.5 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 $1-\delta$,

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

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

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

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

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

Corollary D.2 (Oracle inequality for pseudo-labeling)**.** *Let Assumptions B.1 to B.5 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 D.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 D.1 above to their population version via Lemma D.2.

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

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

On the other hand, the concentration bounds from Lemma D.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. (H.3) since $\tilde{\lambda} > \mu^2/n$. Putting everything together, since by construction $n_{\text{eff}} \le n$, 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 $1-\delta$,

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

□

Remark D.1 (Dependence on the GLM curvature constants)**.** The leading factors in Corollaries D.1 and D.2 are powers of the local condition number K/k of the log-partition function on $[-2C, 2C]$, where $C = RM_2$ bounds the signal. For least squares this ratio is exactly 1 (Table 2), so no such dependence appears. For other GLMs it grows with the signal bound: from Table 2, $K/k \asymp e^{2C}$ for logistic regression and $K/k = e^{4C}$ for Poisson regression, so the constants degrade exponentially in C. This exponential dependence on the signal range is a known feature of curvature-based (self-concordant) analyses of generalized linear models and is not specific to our method; crucially, it affects only the leading constants and leaves the rates in n, n_{eff}, and n_0 unchanged. Our bounds are therefore most informative in the moderate-signal regime $C = O(1)$, which is the practically relevant one once features are standardized and the RKHS norm R is controlled by regularization. We note that a refined analysis based on *generalized self-concordance* (Marteau-Ferey et al., 2019) removes these exponential constants for regularized empirical risk minimization with GLM losses; adapting that machinery to our pseudo-label oracle inequality is a promising direction, although the precise form the resulting bound would take is unclear, and we leave it to future work.

D.3. Proof of the main Theorem B.1

Combining via union bound Corollary D.2 and Lemma D.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

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

where we used again that $\|\boldsymbol{S}_\lambda\| \le r^{-1}k^{-1}$ and $\operatorname{Tr}(\boldsymbol{S}_\lambda) \le 2r^{-1}\operatorname{Tr}((k\boldsymbol{\Sigma}_0+\lambda r^{-1}\boldsymbol{I})^{-1}\boldsymbol{\Sigma}_0)$, 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 $\lambda_1 \ge \frac{\mu^2}{n}\log(1/\delta)\log(n)^{2+\beta/2}$ and $\lambda_p \le \frac{1}{R^4\log(n)^\beta}$.

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:

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

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

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

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

$$\inf_{\lambda\geq\lambda_p}\{A(\lambda)+B(\lambda)\}\geq A\left(\lambda_p\right)\asymp nR^2\mu^2.$$

Meanwhile,

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

Therefore,

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

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 $A(\lambda)\geq A\left(\lambda_j\right)$ and $k\boldsymbol{\Sigma}_0+\mathrm{r}^{-1}\lambda\boldsymbol{I}\preceq\left(\lambda/\lambda_j\right)\left(k\boldsymbol{\Sigma}_0+\mathrm{r}^{-1}\lambda_j\boldsymbol{I}\right)$. Hence,

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

We get

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

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

D.4. Proof of Corollary B.1

Proof. Throughout the proof, we use $\lesssim$ to hide polylogarithmic factors and constants independent of n and n_{eff}. From Theorem B.1, the excess risk is bounded by:

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

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

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

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 $A(\rho)$:

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

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. □

D.5. Optimal $L^2(\mathcal{X};\mathcal{P})$ prediction rate

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

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

E. Proofs of Appendix D

E.1. Proof of Lemma D.1

The key decomposition to use for this lemma comes from the proof of Theorem C.1 in Appendix F.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 C.1 (and intermediate results obtained during its proof), with probability at least $1 - \delta/2$, the following inequalities hold simultaneously:

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

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

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

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

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. (E.1)

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

- For the second term, use Eqs. (E.2) and (E.4):

$$\begin{aligned}
\left\|(\boldsymbol{H}_1^{-1} - \bar{\boldsymbol{H}}_1^{-1})\frac{1}{n_1}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|_Q^2 &= \frac{1}{n_1^2}\left\|\boldsymbol{Q}^{1/2}\bar{\boldsymbol{H}}_1^{-1/2}(\boldsymbol{\Delta}_1(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1})\bar{\boldsymbol{H}}_1^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|^2 \\
&\leq \frac{1}{n_1}\left\|\boldsymbol{Q}^{1/2}\bar{\boldsymbol{H}}_1^{-1/2}\right\|^2 \cdot \left\|(\boldsymbol{\Delta}_1(\boldsymbol{I} + \boldsymbol{\Delta}_1)^{-1})\right\|^2 \cdot \left\|\frac{1}{\sqrt{n_1}}\bar{\boldsymbol{H}}_1^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|^2 \\
&\lesssim \frac{1}{n_1}\left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\| \|\boldsymbol{\Delta}_1\|^2 \sigma^2 \log^2(n_1)\log(8/\delta)\operatorname{Tr}(\hat{\boldsymbol{C}}_{1,\lambda}) \\
&\lesssim \frac{1}{n_1}\left\|\hat{\boldsymbol{T}}_{1,\lambda}\right\| \|\boldsymbol{\Delta}_1\|^2 \sigma^2 \log^2(n_1)\log(8/\delta)M_1^2 \operatorname{Tr}(\boldsymbol{C}_\lambda),
\end{aligned}$$

  where the last inequality uses $\operatorname{Tr}(\hat{\boldsymbol{C}}_{1,\lambda}) \leq 2M_1^2 \operatorname{Tr}(\boldsymbol{C}_\lambda)$.

[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 C.1 when working with dataset $\mathcal{D}_i$.

Recall that $\mathrm{Tr}(\boldsymbol{C}_\lambda) \lesssim \lambda^{-1/2}$ and from Eq. (E.3), we get:

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

so that

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

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

$$\begin{aligned}
\left\|\bar{\boldsymbol{H}}_1^{-1}\frac{1}{n}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\|_Q^2 &= \frac{1}{n_1^2}\left\langle \boldsymbol{\varepsilon}_1, \boldsymbol{X}_1 \bar{\boldsymbol{H}}_1^{-1}\boldsymbol{Q}\bar{\boldsymbol{H}}_1^{-1}\boldsymbol{X}_1^\top \boldsymbol{\varepsilon}_1\right\rangle \\
&\leq \frac{\sigma^2}{n_1^2}\mathrm{Tr}(\boldsymbol{X}_1 \bar{\boldsymbol{H}}_1^{-1}\boldsymbol{Q}\bar{\boldsymbol{H}}_1^{-1}\boldsymbol{X}_1^\top)\log^2(n_1)\log(4/\delta) \\
&\leq \frac{\sigma^2}{n_1}\mathrm{Tr}(\bar{\boldsymbol{H}}_1^{-1}\boldsymbol{Q}\bar{\boldsymbol{H}}_1^{-1}\hat{\boldsymbol{\Sigma}}_1)\log^2(n_1)\log(4/\delta) \\
&= \frac{\sigma^2}{n_1}\mathrm{Tr}(\boldsymbol{Q}\bar{\boldsymbol{H}}_1^{-1}\hat{\boldsymbol{\Sigma}}_1\bar{\boldsymbol{H}}_1^{-1})\log^2(n_1)\log(4/\delta) \\
&\leq \frac{\sigma^2}{kn_1}\mathrm{Tr}(\boldsymbol{Q}\bar{\boldsymbol{H}}_1^{-1})\log^2(n_1)\log(4/\delta) \\
&\leq \frac{\sigma^2}{kn_1}\mathrm{Tr}(\boldsymbol{Q}(k\hat{\boldsymbol{\Sigma}}_1 + \lambda\boldsymbol{I})^{-1})\log^2(n_1)\log(4/\delta) \\
&= \frac{\sigma^2}{kn_1}\mathrm{Tr}(\hat{\boldsymbol{T}}_{1,\lambda})\log^2(n_1)\log(4/\delta).
\end{aligned}$$

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 C.1 now guarantees a $n^{-1/4}$ consistency rate (up to logarithmic factors), otherwise the three inequalities Eqs. (E.1), (E.2) and (E.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

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

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

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 H.2.

E.2. Proof of Lemma D.2

Apply Lemma H.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.

E.3. Proof of Lemma D.3

Start by leveraging the observation made in Eq. (A.2): $R(\hat{f}_\lambda) - R(f^*) \leq \frac{K}{2} \left\| \hat{\boldsymbol{\theta}}_\lambda - \boldsymbol{\theta}^* \right\|_{\Sigma_0}^2$. Then apply Lemma D.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 D.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 $\min_{\lambda \in \Lambda} \leq \min_{\lambda \in \Lambda \cap J}$.

E.4. Proof of Lemma D.4

We want to apply the in-sample oracle inequality from Theorem 4.1 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 D.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 4.1 needed to apply the oracle inequality is verified directly per Assumptions B.2 to B.5, with $C = RM_2$. 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

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

with $\boldsymbol{B} = \boldsymbol{B}_1 + \boldsymbol{B}_2$, $\tilde{\sigma}$ 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 D.1 (for $\boldsymbol{B}_1, \boldsymbol{B}_2, \|\tilde{\varepsilon}\|_{\psi_1}$) and Theorem C.1 (for η).

The variance term:

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

The bias term:

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

Finally, the perturbation term:

$$\frac{K}{n_0 k} \frac{L^2 n_0 \eta^4}{k} \lesssim \frac{L^2 K M_1^6 \sigma^2}{k^2} \frac{\log^6(n_2/\delta)}{n_2}.$$

Combining the three terms together yields the desired result.

E.5. Proof of Lemma D.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 C.1 that gives for any $\epsilon \in (0, 1/2]$

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

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

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

for n large enough.

This exponential tail also yields finite moments such that

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

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.

E.6. Proof of Corollary D.1

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

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

and

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

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 D.5 gives, with high probability,

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

where

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

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

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

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. (E.7) and Eq. (E.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. (E.6), we obtain

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

Therefore,

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

We also compare the out-of-sample excess risk of the selected estimator to its in-sample excess risk. By Eq. (E.6) and Eq. (E.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. (E.7), we get

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

Plugging Eq. (E.9) into Eq. (E.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. (E.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.

F. Proofs of Appendix C.1'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. (A.4), an equivalent formulation of Eq. (5.1) is

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

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 F.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 F.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 B.2.

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:

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

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.

F.1. Coarse-grained analysis of the estimator consistency

Lemma F.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 B.1 to B.5, 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 $E_\rho = \left\{\boldsymbol{\theta} : \|\boldsymbol{S}_{\lambda,\rho}^{1/2}(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq 2\left\|\boldsymbol{S}_{\lambda,\rho}^{-1/2}\nabla L_\lambda(\boldsymbol{\theta}^*)\right\|\right\}$. 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

$$\rho > \frac{2}{C} \max_i \left\{ \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \Phi(\boldsymbol{x}_i) \right\| \cdot \left\| \boldsymbol{S}_{\lambda,\rho}^{-1/2} \nabla L_\lambda(\boldsymbol{\theta}^*) \right\| \right\}, \tag{F.3}$$

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. (F.3), $\hat{\boldsymbol{\theta}}_\lambda \in E_\rho \subset \mathring{\Omega}(\rho)$.

Let us now prove that the inequality (F.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 (F.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 (F.3) for $\rho = 1$

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

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 H.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. (F.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 H.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 H.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:
$$\mathbb{P}\left(\left\langle \varepsilon, \frac{1}{n^2}\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\rangle \lesssim \sigma^2 \log^2(n) \log(4/\delta) \operatorname{Tr}(\frac{1}{n^2}\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top) \;\middle|\; \boldsymbol{X} \right) \geq 1 - \frac{\delta}{2},$$
and thus integrating over the whole event:
$$\mathbb{P}\left(\left\langle \varepsilon, \frac{1}{n^2}\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} \right\rangle \lesssim \sigma^2 \log^2(n) \log(4/\delta) \operatorname{Tr}(\frac{1}{n^2}\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top) \;\middle|\; \mathcal{A} \right) \geq 1 - \frac{\delta}{2}.$$
Now, under $\mathcal{A}$ according to Lemma H.3,
$$\operatorname{Tr}(\frac{1}{n^2}\boldsymbol{X}(k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^\top) = \frac{1}{n}\operatorname{Tr}((k\hat{\boldsymbol{\Sigma}} + \lambda \boldsymbol{I})^{-1}\hat{\boldsymbol{\Sigma}}) \leq \frac{2M_1^2}{n}\operatorname{Tr}((k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}).$$
Then applying again Lemma H.1, we have that
$$\operatorname{Tr}((k\boldsymbol{\Sigma} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Sigma}) = r(\lambda/k)/k \lesssim \lambda^{-1/2\alpha}.$$
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
$$\mathcal{E} := \left\{ RHS \leq cM_1 \lambda^{-1/4\alpha} \left(\lambda^{1/2} \left\| \boldsymbol{\theta}^* \right\| + \sigma M_1 \sqrt{\log(4/\delta)} \lambda^{-1/4\alpha} \frac{\log n}{\sqrt{n}} \right) \right\}$$
has conditional probability
$$\mathbb{P}(\mathcal{E} \mid \mathcal{A}) \geq 1 - \frac{\delta}{2},$$
and
$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{E} \mid \mathcal{A})\, \mathbb{P}(\mathcal{A}) \geq \left(1 - \frac{\delta}{2}\right)^2 \geq 1 - \delta.$$

Crude consistency rate. Therefore, if we set

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

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 (F.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

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

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

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

□

F.2. Proof of Theorem C.1

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

Theorem C.1 (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. (5.1). Let $\beta \in (0, 8)$, and choose $\delta \in (0, 1/e]$. Then, under Assumptions B.1 to B.5, 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 $1-\delta$:

$$\left| \hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0) \right| \leq c M_1 \left(\sigma \log(\frac{n}{\delta}) \frac{\lambda^{-1/4}}{\sqrt{n}} + \lambda^{1/4} \|f^*\|_{\mathcal{F}} \right). \tag{C.5}$$

When setting λ to its lower bound, the result becomes

$$\left| \hat{f}_\lambda(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0) \right| \leq c' \sqrt{M_1^3 \sigma} n^{-1/4} \left(\log(\frac{n}{\delta}) \right)^3. \tag{C.6}$$

Remark F.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

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

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

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

Proof. Recall Eq. (F.2):

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

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 (F.2):

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

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 $\Omega(\rho), \forall \rho \leq 1$. 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 F.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

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

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 F.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 (F.7) separately. Begin by noting the following fact.

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

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

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:*

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

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

[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 H.3, H.1, and Lemma H.2 for the third point, we get the following fact:

Fact F.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:

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

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 (F.7)

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

Computation of the variance term of (F.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:

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

Computation of the perturbation term of (F.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 $\mathcal{B}$:

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

and

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

But by construction from Fact F.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. (F.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 $\frac{cC_1\log n}{\sqrt{n\lambda^{1/\alpha}}}M_1\sigma\sqrt{\log(\frac{8}{\delta})} \leq \frac{1}{2}$,

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

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. (F.10) into Eq. (F.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 $\delta' = \frac{\delta}{2(n+1)} \leq \frac{\delta}{6} \in (0,1)$:

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

and thus

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

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 (F.12) becomes:

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

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.)

□

G. Proof of Section 4's in-sample oracle inequality

Theorem 4.1. *Let $\delta \in (0,1)$. Then there exists a universal constant $C_1 > 0$ such that under Assumption 4.1, the following holds with probability at least $1-\delta$:*

$$\mathcal{R}_{\hat{j}} \leq \inf_{\epsilon > 0}\left\{(1+\epsilon)\mathcal{R}_{j^*} + \frac{C_1 K}{nk}\left(1+\frac{1}{\epsilon}\right)\xi(\tilde{\sigma}, \boldsymbol{B}, \eta, \delta)\right\} \tag{4.4}$$

$$\xi(\tilde{\sigma}, \boldsymbol{B}, \eta, \delta) = K\tilde{\sigma}^2 \log^2(\frac{m}{\delta}) + K\left\|\boldsymbol{B}\right\|^2 + \frac{L^2 n \eta^4}{k}.$$

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

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

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 G.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 G.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 G.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:

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

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 G.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 (G.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:

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

- 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 $\|\tilde{\boldsymbol{\varepsilon}}\|_2^2 \leq 2(n\eta^2 + \|\boldsymbol{B}\|_2^2)$.

  Therefore,

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

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

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

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

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

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

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

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

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

Then using the properties of the energy norm and the boundedness of a'' (Fact G.2) we get:

$$\begin{aligned}
\frac{1}{\sqrt{n}} \left\| \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j \right\|_{\tilde{\boldsymbol{D}}} &\leq \frac{\sqrt{2}}{\sqrt{n}} \left\| \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}_j \right\|_{\bar{\boldsymbol{D}}} \\
&\leq \sqrt{\frac{2}{n}} \left(\left\| \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}^* \right\|_{\bar{\boldsymbol{D}}} + \left\| \boldsymbol{f}_j - \boldsymbol{f}^* \right\|_{\bar{\boldsymbol{D}}} \right) \\
&\leq \sqrt{\frac{2K}{kn}} \left(\left\| \boldsymbol{f}_{\hat{j}} - \boldsymbol{f}^* \right\| + \left\| \boldsymbol{f}_j - \boldsymbol{f}^* \right\| \right) \\
&\leq \sqrt{\frac{K}{k}} \left(\sqrt{\mathcal{R}_j} + \sqrt{\mathcal{R}_{\hat{j}}} \right),
\end{aligned}$$

where the last inequality follows from Eq. (A.3).

Hence, with $\tilde{\xi} = \sqrt{\frac{K}{k}}\xi/\sqrt{n}$, 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. $2ab = 2\frac{a}{\epsilon}\epsilon b \leq \frac{a^2}{\epsilon^2} + \epsilon^2 b^2$, 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. (G.2) to bound $\tilde{\xi}^2$ as:

  $$\tilde{\xi}^2 \lesssim \frac{K}{kn}(K\tilde{\sigma}\log^2(m/\delta) + \|\boldsymbol{B}\|_2^2 (K + 2L^2\eta^2/k) + \frac{L^2 n\eta^4}{k}).$$

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

□

We deduce a bound on the expectation:

Corollary G.1. *Under the same assumptions as Theorem 4.1, we deduce:*

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

Proof. Denoting $W = [\mathcal{R}_{\hat{j}} - (1+\epsilon)\mathcal{R}_{j^*} - \frac{C_1 K}{nk}(1+\frac{1}{\epsilon})(K\left\|\boldsymbol{B}\right\|^2 + L^2 n\eta^4/k)]/\frac{C_1 K}{nk}(1+\frac{1}{\epsilon})K\tilde{\sigma}^2$, we have from the theorem:

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

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

$$\mathbb{P}\left(W \geq t\right) \leq me^{-\sqrt{t}}, \forall t \geq \log^2(m).$$

Integrating to get the expectation yields:

$$\mathbb{E}\left[W\right] = \int_0^\infty \mathbb{P}\left(W > t\right) dt \leq \int_0^{\log^2(m)} 1dt + \int_{\log^2(m)}^\infty me^{-\sqrt{t}}dt = (1+\log(m))^2 + 1.$$

Rearranging the terms and taking the infimum gives the desired result. □

H. Useful lemmas

H.1. Lemmas for the GLM estimator

Lemma H.1. *Let the kernel regularity Assumption B.2 hold, with spectral eigendecay of the form $\mu_j \leq Aj^{-2\alpha}, \alpha \geq 1$, 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 $C > 0$:*

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

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 $\left\langle \Phi(\boldsymbol{x}), (\boldsymbol{\Sigma}+\lambda\boldsymbol{I})^{-1}\Phi(\boldsymbol{x})\right\rangle \leq M_1^2 r(\lambda) \lesssim M_1^2\lambda^{-1/2\alpha}$.

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 $A \neq 1$, use $\lambda = \lambda/A$ in the formula).

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

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

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

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

H.2. Technical lemmas

Lemma H.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 $1-\delta$,*

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

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 $1 - \frac{\delta}{2}$:
$$\tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}} \leq C' \|\tilde{\boldsymbol{x}}\|_{\psi_2}^2 \operatorname{Tr}(\boldsymbol{A}) \log(2/\delta) \lesssim \log^2(n) \operatorname{Tr}(\boldsymbol{A}) \log(2/\delta).$$
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: $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}} + (\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} - \bar{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}}^\top \boldsymbol{A}\bar{\boldsymbol{x}} - \tilde{\boldsymbol{x}}^\top \boldsymbol{A}\tilde{\boldsymbol{x}}) \lesssim \|\boldsymbol{A}\| \left(n^{1-C} + n^{1-C/2}\log n\right) + \log^2 n \operatorname{Tr}(\boldsymbol{A}) \log(2/\delta)$.

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 H.3. *Let Assumptions B.1 and B.2 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,

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

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.

□

H.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. (B.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

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

and

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

On the other hand, we have simply:

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

hence:

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

I. Complement to the numerical experiments and ablation study

I.1. Implementation details

We optimize the ridge-regularized kernel GLM Eq. (5.1) in function space $\mathbf{f} = \boldsymbol{K}\boldsymbol{\alpha}$ via Fisher scoring (McCullagh & 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 $V_i = \mathrm{Var}[y_i \mid \eta_i]$, we compute the standard GLM weights $\boldsymbol{W} = \mathrm{diag}((\frac{d\mu_i}{d\eta_i})^2/V_i)$ 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. (A.5).

To ensure stable convergence of our conjugate gradient (Hestenes & 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 $\mathbf{f} = \boldsymbol{S}^{-1}\boldsymbol{u}$.

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://anonymous.4open.science/r/pseudo-label-validation-uda-2601`.

I.2. 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 $y_i \mid \boldsymbol{x}_i \sim \mathrm{Bernoulli}(\sigma(f^*(\boldsymbol{x}_i)))$. We can thus denote the conditional mean as a probability $p_i^* := \sigma(f^*(\boldsymbol{x}_i))$. Equivalently, $f^*(\boldsymbol{x}_i) = \log(\frac{p_i^*}{1-p_i^*}) = \mathrm{logit}(p_i^*)$. 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. (2.3) recovers the usual log-loss with "soft labels" or cross-entropy between the true distribution p^* and the estimate p:

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

For simulation experiments, our feature space $\mathcal{X}$ is the unit interval $[0, 1]$, and the kernel K is the first-order Sobolev kernel $K(u, v) = \min\{u, v\}$. 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 $n \geq 2$, 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 $y_i \mid x_i \sim \mathrm{Bernoulli}\left(\sigma(f^*(x_i))\right)$, 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 coefficient in front of the $\cos$ in f^* reflects the fact that the decision boundary for the logistic regression is crucial for training, so the source must contain enough points in the region around 0 to learn decent pseudo-labels. The quantity B controls the strength of the covariate shift. A larger sample size allows for a larger shift. We chose a relatively large covariate shift to better illustrate the relative performance of our method.

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

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

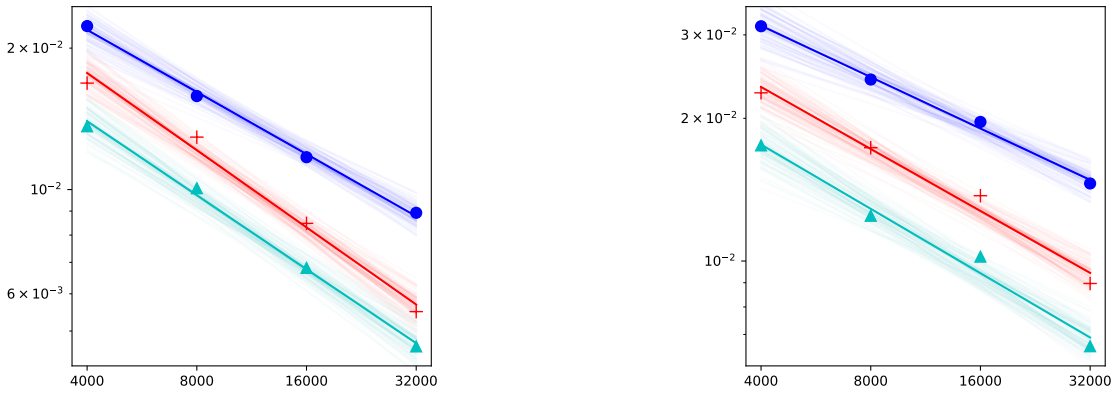
on $\mathcal{D}_2$. The latter is the standard hold-out validation method that ignores the covariate shift.

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. (A.1).

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 I.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 vertical ordering of the curves allows to compare the target excess risks of the selected estimators: the oracle is lowest, as expected, since it selects λ using the true target regression function; pseudo-labeling closely tracks this oracle benchmark; 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 $O(n^{-\beta})$ for some $\beta > 0$. Linear fits to the log-transformed data give $\beta = 0.546(0.047)$, $0.523(0.051)$, 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. To visualize the estimator stability, we have included thin shaded regions representing the linear fits of 100 bootstrap replicates.

Figure 1b shows the results for the same experiments, with covariate shift strength $B = n^{0.45}$. Estimates of the β coefficients are $0.434(0.049)$, $0.451(0.053)$ and $0.360(0.047)$ for the pseudo-labeling, oracle, and naive methods, respectively.



(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.

I.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 4.1) 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. 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. Consistent with the asymmetric trade-off, the excess risk decreases sharply as $\tilde{\lambda}$ is reduced from the baseline and then *flattens* once $\tilde{\lambda}$ reaches the lower end of the range admissible under Theorem C.1: pushing the penalty further yields no additional benefit, since below this floor the imputer's variance, rather than its bias, dominates its pointwise error. The location of this floor grows with $1/n_2$, so in smaller-sample regimes the optimum is reached at a larger $\tilde{\lambda}$ and controlling the imputer's variance becomes correspondingly more important (cf. Appendix I.4).

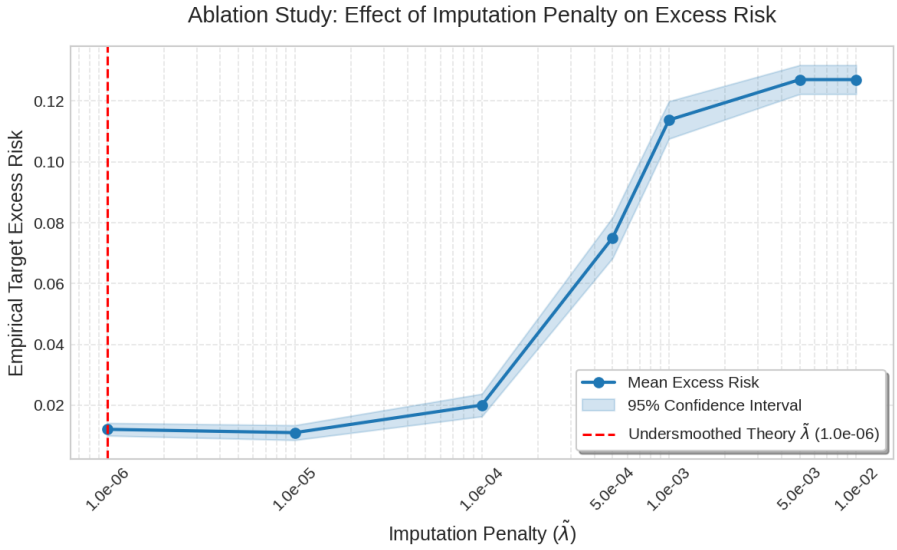


Figure 2. Pseudo-labeling final estimator's excess risk as a function of the imputer's penalty.

Again, full details and reproducible code can be found at https://anonymous.4open.science/r/pseudo-label-validation-uda-2601.

I.4. Details about the real data experiments

The Raisin dataset[7] (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 sub-sampling 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 $l = 3$.

To first realize 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 plots the results for four different metrics. We can observe that only a specific range of regularization (between 10^{-3} and 10^{-2})

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

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.

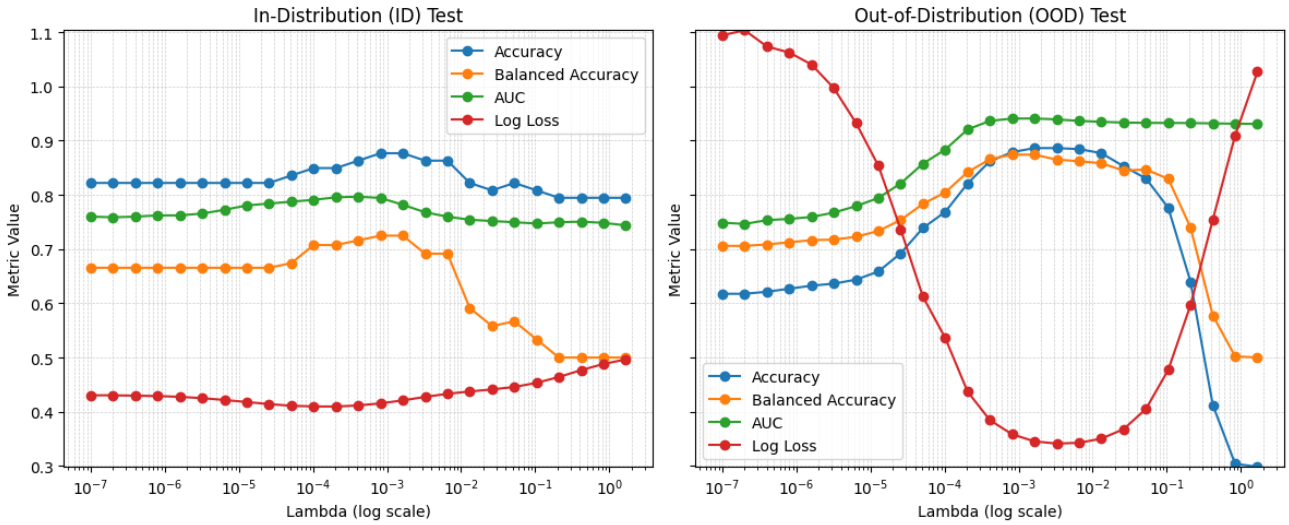


Figure 3. Baseline logistic regression performance on ID and OOD test data, when trained only on ID data, as a function of the ridge regularization parameter λ.

We realize our pseudo-labeling experiment in the same fashion as for the synthetic data in the previous paragraph, by comparing it to a naive method that selects the best model on source validation data, and 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.

We run it for 100 different random seeds for significance and report the results in Table 1. While the underlying task in this example is binary classification, our primary objective is the regression of the conditional probability function. Consequently, we evaluate performance using the empirical risk (equivalently, the log-loss), as it is a strictly proper scoring rule that captures the full calibration of the predicted probabilities. 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 & 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. While our theoretical analysis emphasizes the primacy of low bias for the imputation model, controlling variance proves critical as well in the small-sample regime to prevent performance degradation. Empirically, our pseudo-labeling approach consistently outperforms the other baselines by effectively adapting to the underlying target distribution. In particular, the higher variance of the importance-weighted methods reflects the practical instability of density-ratio estimation in finite samples.

Let us explain in more details the experiments. 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, $\mathcal{D}_0 = \mathcal{D}_{0,\text{sel}} \cup \mathcal{D}_{0,\text{test}}$, and report all final metrics on $\mathcal{D}_{0,\text{test}}$.

Table 3. Risk mean estimate across seeds with confidence interval (CI) and standard error (SE).

Method	Mean	95% CI	SE
Naive	0.442	[0.414, 0.469]	0.014
KMM + IW-CV	0.449	[0.431, 0.468]	0.009
KLIEP + IW-CV	0.434	[0.412, 0.465]	0.014
Pseudo-labeling	0.383	[0.372, 0.394]	0.006
Oracle	0.376	[0.362, 0.389]	0.007

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 before, due to the small sample size, we use a slightly modified selection procedure compared to Algorithm 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.

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. Then we ran Algorithm 2 over 100 random seeds. In Section 5.1 we report results for $K = 2, R = 6$, our default: with two folds each split allocates half the source data to the candidates and half to the imputer, matching the $n_1 = n_2 = n/2$ split of Algorithm 1; the $R = 6$ repetitions only reduce the variance of the risk estimates and do not change this ratio. Table 3 reports the same experiment with $K = 5, R = 2$, which we regard as an *ablation on the imputer's data share* rather than an alternative method: with five folds the imputer is trained on $4/5$ of the data instead of $1/2$. Consistent with our bias-driven analysis (Section 4), the better-resourced imputer has lower bias and the pseudo-labeling gap to the oracle shrinks accordingly (from 0.055 to 0.007), while the oracle's own risk is essentially unchanged. The improvement is therefore attributable to the imputer, as the theory predicts, not to the selection protocol per se.

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 (I.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.

I.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. We use hospitals 0–3 as the labeled source domain and hospital 4 as the unlabeled target domain, following the standard WILDS (Koh et al., 2021) evaluation protocol. The dataset contains 302,436 source training patches, 33,560 source validation patches, and 85,054 target 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 (lr=1e-5, weight decay=0.01, cosine annealing schedule, batch size=256). The epoch number $t \in \{1, ..., 9\}$ serves as the regularization parameter: early epochs correspond to strong regularization (small deviation from pretrained weights), late epochs to weak regularization (full source adaptation). 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

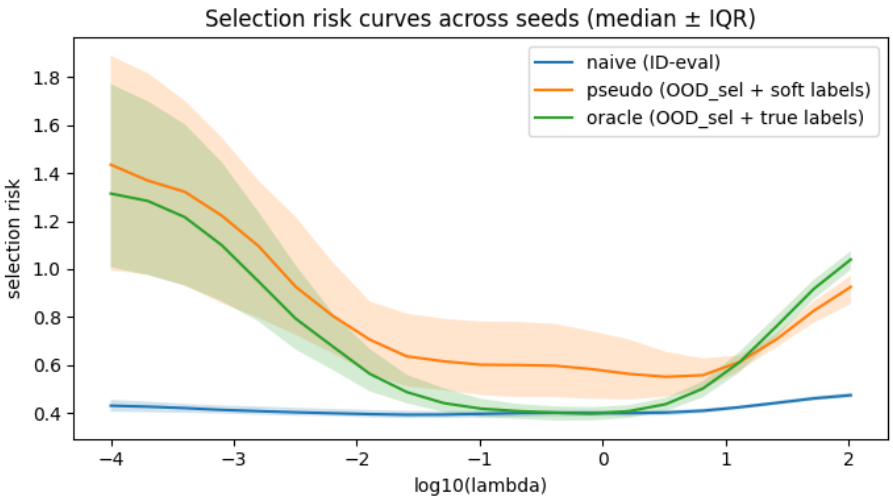


Figure 4. Selection risk curves of the three methods.

Table 4. Early stopping experiment: selected epoch by each method and performance metrics.

Method	Epoch	Target NLL	Target accuracy
Naive	8	1.942	0.531
Pseudo-labeling	1	0.655	0.732
Oracle	1	0.655	0.732

are generated once on a target selection set ($\mathcal{D}_0$, 50% of hospital 4, sampled with seed 42) 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 $\mathcal{D}_0$), and oracle (true target NLL on $\mathcal{D}_0$, not available in practice). The method with the lowest selection signal selects its corresponding checkpoint. Final evaluation is performed on the held-out target test set ($\mathcal{D}_{\text{test}}$, remaining 50% of hospital 4), which is never used during selection.

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 relevant Gram matrix scales as n_0^2 (or as the source–target product $n \times n_0$ for the cross-kernel), i.e. on the order of 10^9 entries—several gigabytes in single precision—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. The oracle and pseudo-label selection losses increase monotonically from epoch 1, correctly identifying that fine-tuning degrades target performance immediately. The naive validation loss decreases monotonically as the model fits source data better, reaching its minimum at epoch 8 — by which point the model has catastrophically overfit to source-specific features. The pseudo-labeling method faithfully tracks the oracle throughout, enabling correct early stopping selection without access to target labels. Table 4 shows the selected epochs and associated target test performance accuracy and negative log-likelihood (NLL).

Scope and limitations of this experiment. On Camelyon17 both the oracle and pseudo-target risks are *monotone* in the epoch (Table 4), we therefore regard this experiment as a real-world *demonstration* that the surrogate loss tracks target risk on a large-scale deep model, rather than as a discriminating test of the selection rule; the harder, non-monotone test is the Raisin study of Appendix I.4, where naive and importance-weighted selection genuinely underperform. Second, the imputer is itself a single DINOv2 fine-tuned on $\mathcal{D}_2$ (hospitals 0–3), which shares the source distribution of the candidates.

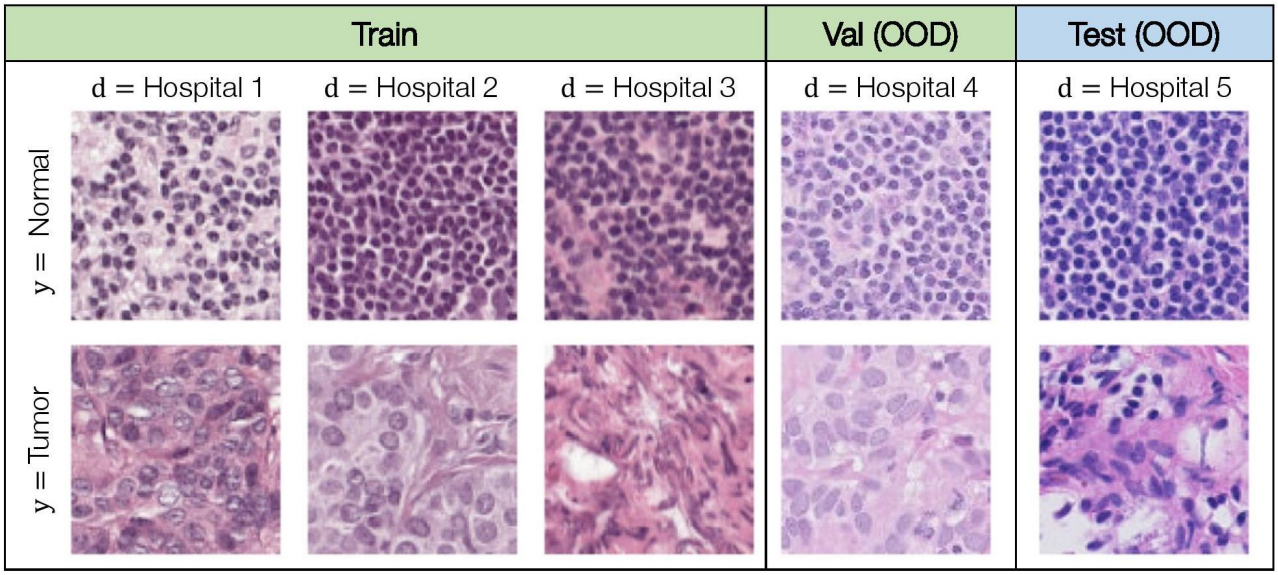


Figure 5. Camelyon17 dataset from Koh et al. (2021)

We note that our analysis requires that the imputer's pseudo-labels *rank* the candidates as the oracle would, which only needs the pointwise-consistency condition of Assumption 4.1. This condition is verified for kernel GLMs (Theorem C.1) but remains an empirical assumption for the deep imputer; we use a single fine-tuned model rather than an ensemble, consistent with the undersmoothing principle of Section 4. A direct check—comparing the imputer's pseudo-labels against held-out hospital-4 ground truth on a small labeled subset—is a natural diagnostic of this assumption. Third, the present comparison on Camelyon17 is against naive source validation alone. While kernel density-ratio validation (KMM/KLIEP) is intractable in this regime (above), discriminator-based density-ratio validation such as DEV (You et al., 2019) and confidence- or entropy-based criteria such as SND (Saito et al., 2021) remain applicable to deep models and are the appropriate target-label-free baselines here; a systematic comparison against them would more firmly anchor the claim that the surrogate loss tracks target risk in the deep setting.

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

1: **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 $\mathcal{D}_{0,\mathrm{sel}}$ and $\mathcal{D}_{0,\mathrm{test}}$. 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.

2: **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):

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

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

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

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

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

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

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

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

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

8: **Step 4 (Selection).** Select regularization parameters independently:

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

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

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

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

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