

Coupled Training with Privileged Information and Unlabeled Data

Jiahao Shi¹ Omar Hagrass² Jason M. Klusowski²

Abstract

In many prediction problems, we have extra information during training (for example, measurements that are expensive or slow to collect) that will not be available when the model is deployed. A common strategy is to first train a model that uses all training information, then use its predictions on unlabeled examples to train a second model that only uses the inputs available at test time. However, when the extra training-only information is weak or noisy, this Two-Stage approach can mislead the deployment model and even hurt accuracy. We propose a joint training method that learns the two models together, so the deployment model can benefit from the extra information only when it actually helps, instead of inheriting its mistakes. We provide guarantees that describe when joint training improves prediction accuracy and analyze a simple alternating training algorithm for large, high-dimensional models. Experiments on synthetic data and real-world prediction tasks show that our approach avoids these failures and robustly outperforms standard Two-Stage baselines.

1. Introduction

Many modern learning settings provide extra information during training that cannot be used at deployment. Alongside the primary features X (e.g., an image or routine clinical measurements) and responses Y (e.g., a continuous outcome such as disease severity or a lab value), training examples may also include privileged features W (e.g., expensive biomarkers, specialist assessments). Because W may be unavailable, costly, or impractical to obtain at test time, the deployed predictor must depend only on X . At the same time, a large fraction of the training data may contain W but not the true labels. The goal is therefore to use W to

¹Department of Electrical and Computer Engineering, Princeton University ²Department of Operations Research and Financial Engineering, Princeton University. Correspondence to: Jiahao Shi <js6028@princeton.edu>.

improve training while still producing a final predictor that uses only X . This setting is motivated by many applications, including medical domains where reliable labels are expensive (Rajkomar et al., 2019; Fries et al., 2019), forecasting and longitudinal studies where intermediate or future measurements are observed only during training (Hogan et al., 2004), as well as transfer learning (Zhuang et al., 2020) and distribution shift settings (Koh et al., 2021).

Problem Formulation. We formalize this setting as follows. Let $\mathcal{Z} = \mathcal{X} \times \mathcal{W}$ and define $Z = (X, W) \in \mathcal{Z}$. We observe n labeled samples

$$\mathcal{D}_L = \{(Z_i, Y_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{X,W,Y},$$

and $m := N - n \gg n$ unlabeled samples

$$\mathcal{D}_U = \{Z_j\}_{j=n+1}^N \stackrel{\text{i.i.d.}}{\sim} P_{X,W},$$

where $P_{X,W,Y}$ is a distribution on $\mathcal{X} \times \mathcal{W} \times \mathbb{R}$ and $P_{X,W}$ is its marginal over $\mathcal{Z}$.

Our goal is to estimate a prediction rule that minimizes the expected risk

$$\mathbb{E}[\ell(Y, f(X))],$$

where ℓ is a general loss function. When ℓ is the square loss, i.e., $\ell(y, y') = (y - y')^2$, the regression function

$$\mu(x) := \mathbb{E}[Y \mid X = x],$$

uniquely minimizes the squared risk $\mathbb{E}[(Y - f(X))^2]$ over measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$, assuming $\mathbb{E}[Y^2] < \infty$. Given this setup, the central question is whether we can use the privileged information W to achieve better performance than training on X alone, while still ensuring that the deployed predictor depends only on X .

Two-Stage Procedure. One common way to leverage privileged data is through a Two-Stage procedure. In the first stage, we estimate a pseudo-response function $\hat{g}$, which we refer to as the *rich-view model*, using all available labeled samples $\{(Z_i, Y_i)\}_{i=1}^n$ by minimizing the empirical risk

$$\hat{g} \in \operatorname{argmin}_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(Z_i)),$$

where $\mathcal{G}$ is a collection of functions defined on $\mathcal{Z}$. The fitted function $\hat{g}$ is then used to transfer knowledge by imputing

pseudo-responses $\hat{Y}_j = \hat{g}(Z_j)$ (or $\hat{Y}_j = \mathbf{1}(\hat{g}(Z_j) \geq 1/2)$ for 0/1 binary classification) for the unlabeled samples $\{Z_j\}_{j=n+1}^N$. In the second stage, we estimate the target function f , which we refer to as the *deployment model*, from the combined dataset $\{(X_i, Y_i)\}_{i=1}^n \cup \{(X_j, \hat{Y}_j)\}_{j=n+1}^N$ by minimizing

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \left(\sum_{i=1}^n \ell(Y_i, f(X_i)) + \sum_{j=n+1}^N \ell(\hat{Y}_j, f(X_j)) \right),$$

where $\mathcal{F}$ is a collection of functions defined on $\mathcal{X}$. This general strategy was studied in Hou et al. (2023) for specific parametric models and extended to more general nonparametric settings in Xia & Wainwright (2024).

Although it is appealingly simple, this Two-Stage pseudo-labeling strategy can be fragile when the privileged features W provide only weak or noisy information about the response. In such regimes, the first-stage estimator $\hat{g}$ may produce pseudo-responses that are poorly aligned with the true regression function μ , and these errors are then propagated to the second stage through the imputed labels. As a consequence, the deployment model $\hat{f}$ may be biased toward artifacts introduced by $\hat{g}$, leading to degraded performance relative to training on the labeled data alone. This phenomenon, often referred to as *negative transfer*, is especially pronounced when the predictive advantage of (X, W) over X is marginal, or when the privileged variables contain high-dimensional nuisance components that overwhelm their useful signal. In such settings, naive Two-Stage procedures can perform strictly worse than supervised learning using only the labeled samples, despite access to abundant unlabeled data (Xia & Wainwright, 2024). These limitations motivate approaches that more carefully control the influence of privileged information and adaptively balance the contribution of W during training.

Our Contributions. We summarize our main contributions as follows.

- **Coupled Training Framework.** We propose a unified estimation framework that jointly learns the deployment and rich-view models. By flexibly calibrating the influence of privileged information, our approach effectively mitigates the negative transfer often observed in Two-Stage pseudo-labeling methods.
- **Efficient Algorithms.** We develop efficient implementations of the coupled training framework and extend it to high-dimensional settings using a greedy forward selection procedure with provable optimization guarantees.

- **Theoretical and Empirical Analysis.** We provide theoretical insights into the conditions under which unlabeled data improves estimation and demonstrate, through synthetic experiments and real-world regression and classification benchmarks, that our method consistently outperforms standard baselines when privileged signals are noisy or weak.

1.1. Pseudo-Response Calibration with Coupled Training

The main limitation of the Two-Stage approach is that the pseudo-responses produced by the rich-view model are treated as fixed targets in the second stage, with no mechanism to correct or attenuate their influence when the privileged signal is weak or noisy. To address this issue, we propose a coupled training procedure that jointly estimates the deployment model f and the rich-view model g through the updates below. The core idea is to allow information to flow in both directions, with g providing pseudo-responses that enrich the effective sample size for learning f , while the current estimate of f is in turn used to recalibrate g on the unlabeled data, subject to an explicit constraint that limits deviation from the labeled responses. This coupling yields a flexible regularization of the privileged information and provides a principled mechanism for avoiding negative transfer.

We now describe the resulting algorithm and its associated optimization formulation. For simplicity, we focus on regression, but the same ideas can be adapted for classification, e.g., by treating $\hat{Y}$ as soft labels and using logistic loss.

Coupled Training Algorithm. Initialize with any feasible $g_0 \in \mathcal{G}$ and fix a constraint level $\nu \geq 0$. For $k = 1, 2, \dots$, alternate until convergence.

1. **Update f .** Given the current pseudo-responses $\hat{Y}_j = g_{k-1}(Z_j)$, update f by solving

$$f_k \in \operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{N} \left(\sum_{i=1}^n \ell(Y_i, f(X_i)) + \sum_{j=n+1}^N \ell(g_{k-1}(Z_j), f(X_j)) \right).$$

2. **Recalibrate g .** Given the current deployment model f_k , update g by solving

$$g_k \in \operatorname{argmin}_{g \in \mathcal{G}} \frac{1}{m} \sum_{j=n+1}^N \ell(g(Z_j), f_k(X_j))$$

$$\text{s.t. } \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(Z_i)) \leq \nu.$$

If $\mathcal{F} \subset L^2(\mathcal{X})$ and $\mathcal{G} \subset L^2(\mathcal{Z})$ are convex function classes, the loss $\ell(y, y')$ is jointly convex in its arguments (y, y') (e.g., square loss) and the constraint set is feasible, then the

coupled training algorithm gives exact updates for the joint convex optimization problem.

$$\begin{aligned} \min_{(f,g) \in \mathcal{F} \times \mathcal{G}} \quad & \frac{1}{N} \left(\sum_{i=1}^n \ell(Y_i, f(X_i)) \right. \\ & \left. + \sum_{j=n+1}^N \ell(g(Z_j), f(X_j)) \right) \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^n \ell(Y_i, g(Z_i)) \leq \nu. \end{aligned} \quad (1)$$

In particular, the objective value is nonincreasing along the iterates. Under standard regularity conditions for these exact updates, every cluster point is a stationary point, and hence globally optimal by convexity for (1); see [Grippe & Sciadrone \(2000\)](#).

The implementation of the coupled training algorithm depends on the model class. For linear models with square loss, the penalized joint objective is quadratic in the coefficients of (f, g) , so its first-order conditions reduce to a single block linear system. Thus, we solve this system in closed form. For differentiable nonlinear models, such closed-form equations are unavailable, so we compute the same coupled training updates using gradient-based solvers. For tree ensembles, we use the penalized Lagrangian form in (2), so each alternating update reduces to a standard weighted tree regression. For high-dimensional dictionary models, the updates are computed using the greedy selection routine in Section 3. See Appendix B for implementation details. Throughout, “joint” training means that f and g are coupled through one objective, not that every experiment uses a single simultaneous optimizer over all parameters. Henceforth, we restrict attention to the square loss $\ell(y, y') = (y - y')^2$, which allows for a tractable analysis. Of course, the proposed coupled training algorithm is not restricted to this choice of loss.

We identify any $f : \mathcal{X} \rightarrow \mathbb{R}$ with its lift $\tilde{f} : \mathcal{Z} \rightarrow \mathbb{R}$, defined by $\tilde{f}(x, w) = f(x)$; below, we simply write f for this lift when no confusion is possible.

Negative Transfer. The constraint level ν governs how much the rich-view model g is allowed to deviate from the labeled responses, and therefore how strongly privileged information can influence the deployment model f . In particular, feasibility requires

$$\nu \geq \min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^n (Y_i - g(Z_i))^2.$$

If $\mathcal{F} \subseteq \mathcal{G}$ (identifying $f \in \mathcal{F}$ with its lift $\tilde{f}(x, w) = f(x)$), then choosing ν large enough that the labeled least-squares fit over $\mathcal{F}$ is feasible with $g = f$ makes the unlabeled agreement term vanish, and the solution reduces to the labeled least-squares fit over $\mathcal{F}$. In contrast, choosing ν close to the

minimum labeled risk over $\mathcal{G}$ forces g to be (approximately) the labeled least-squares fit over $\mathcal{G}$, yielding behavior akin to the aforementioned Two-Stage pseudo-labeling procedure. Thus, choosing ν too small can overemphasize noisy pseudo-responses and cause negative transfer, while choosing larger ν attenuates the influence of g and interpolates back toward supervised learning that ignores W .

To illustrate this behavior, we consider a linear model

$$Y = \beta^\top X + \theta^\top W + \varepsilon,$$

where (X, W) are jointly Gaussian and ε is independent mean-zero Gaussian noise. The strength of the privileged feature is governed by $\|\theta\|_2$, while all other aspects of the data-generating process are held fixed.

Figure 1 illustrates the performance of the proposed method under varying levels of privileged signal strength. In panel (a), where $\|\theta\|_2$ is large, the privileged variables are highly informative, and the optimal performance is achieved for small values of ν . In this regime, the method behaves like a Two-Stage procedure, leveraging the strong privileged signal to improve estimation accuracy.

In contrast, panel (b) corresponds to a setting in which the privileged signal is weak. Here, aggressively incorporating pseudo-responses can degrade performance, yielding worse error than the least-squares estimator (OLS) based only on the labeled pairs $\{(X_i, Y_i)\}_{i=1}^n$. This phenomenon of negative transfer was also noted in [Xia & Wainwright \(2024, Section 2.4.2\)](#) as a limitation of the Two-Stage approach. Together, these results show that the proposed coupled training framework adaptively moderates the influence of the rich-view model, interpolating between regimes in which privileged information helps or hurts prediction, and thereby yields a robust, principled approach to leveraging privileged information.

1.2. Related Work

We place our coupled training framework in context by reviewing connections to learning using privileged information, semi-supervised learning (especially pseudo-labeling), multi-view agreement methods, and related optimization approaches.

Learning Using Privileged Information. Our setup is closely related to Learning Using Privileged Information (LUPI) ([Vapnik & Vashist, 2009](#); [Vapnik et al., 2015](#)), where privileged variables W are available during training but the goal is to learn a predictor that depends only on X . However, in contrast to our framework, classical LUPI formulations assume fully labeled data and do not leverage unlabeled samples.

Semi-Supervised Learning. Learning from both labeled and unlabeled data has been widely studied across statistics and machine learning under the name semi-supervised learning (SSL) ([Chapelle et al., 2006](#)).

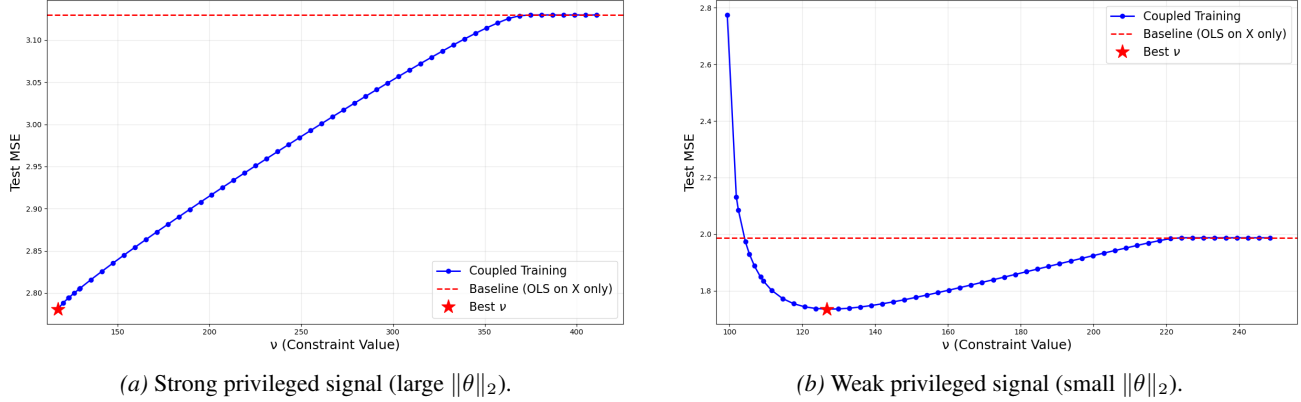


Figure 1. **Linear Gaussian signal strength.** Performance of the proposed method under varying levels of privileged signal strength.

Among SSL methods, pseudo-labeling is a particularly prominent approach. In classification, pseudo-labeling was explored in the context of support vector machines by Joachims (1999) and later extended to deep learning frameworks by Lee (2013). In regression, pseudo-responses were constructed via kernel smoothing and imputation in linear models (Chakraborty & Cai, 2018), extended to kernel ridge regression (Wang, 2023), and studied in general settings by Xia & Wainwright (2024). From an inferential perspective, mean estimation with SSL data was investigated by Zhang et al. (2019); Zhang & Bradic (2021), while Laine & Aila (2017) proposed an ensemble-based method that iteratively assigns pseudo-responses using historical predictions. A parallel line of work is weakly supervised learning, which relies on weak or noisy labels rather than ground truth. The data programming paradigm of Ratner et al. (2016; 2017) assumes access to labeling functions and was applied to MRI analysis by Fries et al. (2019). Weak observations were also used for data augmentation by Robinson et al. (2020), who established theoretical guarantees under universal central conditions. Closely related is the literature on learning with noisy labels (Natarajan et al., 2013; Song et al., 2022).

In contrast to these approaches, which primarily rely on pseudo-responses or weak supervision, our framework explicitly controls the influence of privileged information within an adaptive self-training scheme.

Multi-View Learning. The objective in (1) is related to the *co-regularized least-squares* objective of Sindhwani et al. (2005), which encourages predictors built from different views to agree on unlabeled data. In their multi-view SSL setting, one observes labeled examples $\{(X_i, Y_i)\}_{i=1}^n$ and unlabeled examples $\{X_i\}_{i=n+1}^N$, where each input $X_i = (X_i^{(1)}, X_i^{(2)})$ comes with two complementary representations $X_i^{(1)} \in \mathcal{X}^{(1)}$ and $X_i^{(2)} \in \mathcal{X}^{(2)}$. The goal is to learn $h = (h^{(1)}, h^{(2)})$ so that each component fits the labeled data in its own view, while their predictions are

pushed to match on the unlabeled inputs (see Sindhwani et al. (2005, Section 3)). Related agreement-style objectives also appear in Ding et al. (2022), though their setting is fully supervised and does not use unlabeled data.

In our case, (1) uses a similar kind of agreement term, but for a different reason. We work in a *privileged information* setting where W is available during training but not at deployment, so only $f(X)$ can be used at test time. That makes the coupling inherently asymmetric, with f as the deployment model and g as an auxiliary rich-view model used to pass information to f during training. We also take g to act on the joint input $Z = (X, W)$, rather than on W alone. This lets g capture effects of the privileged variables that depend on X (including X - W interactions) and helps translate whatever signal is present in W into improvements in f .

Greedy and Dictionary-Based Optimization. In Section 3, we restrict the deployment and rich-view models to dictionary spans and fit them using an *alternating forward selection* procedure. This connects our method to the literature on greedy approximation in Hilbert spaces, where algorithms construct an additive predictor by sequentially selecting atoms that best reduce a squared-loss objective and enjoy deterministic approximation guarantees with sublinear objective or approximation error decay (Barron et al., 2008; DeVore & Temlyakov, 1996). It can also be viewed as a block-coordinate variant of classical forward (stepwise) selection in regression (Hastie et al., 2009, Sec. 3.3.2).

Our setting differs in that we perform greedy atom selection within each block while alternating between updating f with g fixed and updating g with f fixed, so the greedy search is carried out in a product space with interdependent residuals. Although support selection is combinatorial, we prove that the alternating greedy iterates achieve a global sublinear decay of the empirical coupled objective (Theorem 3.1), extending classical greedy approximation analyses to the coupled privileged information setting. This is comple-

mentary to blockwise optimization analyses for nonconvex problems such as dictionary learning, which typically focus on reconstruction or coefficient recovery under structural assumptions (Chatterji & Bartlett, 2017; Agarwal et al., 2014; Ruetz & Schnass, 2023); here the dictionaries are fixed and our guarantee controls the empirical comparison term through an optimization error bound.

1.3. Notation

Let $\mathcal{F} \subset L^2(\mathcal{X})$ and $\mathcal{G} \subset L^2(\mathcal{Z})$ be closed, convex function classes. For a nonempty closed convex set $\mathcal{C}$ in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$, let $\Pi_{\mathcal{C}} h \in \arg \min_{c \in \mathcal{C}} \|h - c\|_{\mathcal{H}}^2$ denote the (metric) projection. When $\mathcal{C}$ is a closed linear subspace, $\Pi_{\mathcal{C}}$ coincides with the orthogonal projector.

For fixed points $x^N = (x_1, \dots, x_N) \in \mathcal{X}^N$ and $z^N = (z_1, \dots, z_N) \in \mathcal{Z}^N$, where $z_i = (x_i, w_i)$, let $\hat{P}_N^X$ and $\hat{P}_N^Z$ denote the corresponding normalized empirical measures.

For functions $f : \mathcal{X} \rightarrow \mathbb{R}$ and $g : \mathcal{Z} \rightarrow \mathbb{R}$, we write $\|f\|_{L^p(\hat{P}_N^X)} := \left(\frac{1}{N} \sum_{i=1}^N |f(x_i)|^p \right)^{1/p}$, and $\|g\|_{L^p(\hat{P}_N^Z)} := \left(\frac{1}{N} \sum_{i=1}^N |g(z_i)|^p \right)^{1/p}$. For a class $\mathcal{F} \times \mathcal{G}$, define the max norm $\|(f, g)\|_{1,\infty} = \max\{\|f\|_{L^1(\hat{P}_N^X)}, \|g\|_{L^1(\hat{P}_N^Z)}\}$. An ϵ -cover of $\mathcal{F} \times \mathcal{G}$ with respect to $\|\cdot\|_{1,\infty}$ is a finite collection $\{(f_1, g_1), \dots, (f_M, g_M)\} \subset \mathcal{F} \times \mathcal{G}$ such that every $(f, g) \in \mathcal{F} \times \mathcal{G}$ is within ϵ of some (f_j, g_j) under this norm, and the minimal such M is denoted by $\mathcal{N}_{1,\infty}(\epsilon, \mathcal{F} \times \mathcal{G}, z^N)$.

2. Risk Bounds

Lagrangian Relaxation. The constrained problem (1) admits a Lagrangian relaxation, leading to the penalized objective

$$\begin{aligned} \hat{\mathcal{L}}(f, g; \lambda) := & \frac{1}{N} \left(\sum_{i \leq n} (Y_i - f(X_i))^2 \right. \\ & \left. + \sum_{j > n} (g(Z_j) - f(X_j))^2 + \lambda \sum_{i \leq n} (Y_i - g(Z_i))^2 \right). \end{aligned} \quad (2)$$

Under the usual convex duality conditions for (1), for example feasibility together with a suitable constraint qualification, the constrained problem admits KKT multipliers. In that case, for a primal solution $(\hat{f}, \hat{g})$ there exists a multiplier $\tilde{\lambda} \geq 0$ such that $(\hat{f}, \hat{g})$ minimizes the penalized objective (2) with $\lambda = (N/n)\tilde{\lambda}$ and satisfies the complementary-slackness condition

$$\tilde{\lambda} \left(\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{g}(Z_i))^2 - \nu \right) = 0.$$

In the rest of the paper we treat the rescaled parameter λ as a tuning parameter and work directly with the penalized formulation.

Note that λ plays a role analogous to ν , but in the opposite direction. Smaller λ relaxes the effective constraint on g and

attenuates the influence of pseudo-responses, while larger λ forces g to fit the labeled data more closely and increases the impact of pseudo-labeling. In particular, if $\mathcal{F} \subseteq \mathcal{G}$, then as $\lambda \rightarrow 0$, the unlabeled data play no role and the solution reduces to the ordinary least-squares fit based only on the labeled data. As $\lambda \rightarrow \infty$, the solution approaches the Two-Stage procedure. In this case, since $\hat{\mathcal{L}}(\hat{f}, \hat{g}; \lambda) \leq \hat{\mathcal{L}}(\hat{f}, \hat{f}; \lambda)$, we also have the bound $\frac{1}{m} \sum_{j=n+1}^N (\hat{Y}_j - \hat{f}(X_j))^2 \leq \frac{\lambda n}{m} \cdot \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{f}(X_i))^2$, meaning that the (pseudo-) training error of $\hat{f}$ on the unlabeled features is at most $\lambda n/m$ times the training error of $\hat{f}$ on the labeled features.

Population Minimizers. We now turn to a theoretical analysis of the proposed method and derive bounds on the prediction error of the learned deployment model $\hat{f}$. To this end, we consider the population-level counterpart of the empirical objective in (2), given by

$$\begin{aligned} \mathcal{L}(f, g; \lambda) := & \frac{n}{N} \mathbb{E}[(Y - f(X))^2] \\ & + \frac{m}{N} \mathbb{E}[(g(Z) - f(X))^2] \\ & + \lambda \frac{n}{N} \mathbb{E}[(Y - g(Z))^2]. \end{aligned} \quad (3)$$

Recall we identify any $f : \mathcal{X} \rightarrow \mathbb{R}$ with its lift to $\mathcal{Z}$ via $\tilde{f}(x, w) = f(x)$. We define the rich-view regression function

$$\eta(z) := \mathbb{E}[Y \mid Z = z],$$

which is related to μ via $\mu(x) = \mathbb{E}[\eta(Z) \mid X = x]$.

Theorem 2.1 (Population Minimizers). *Any $(f^*, g^*) \in \mathcal{F} \times \mathcal{G}$ minimizing (3) satisfies the fixed point equations.*

$$\begin{aligned} f^* &= \Pi_{\mathcal{F}} \left(\frac{n}{N} \mu + \frac{m}{N} \mathbb{E}[g^*(Z) \mid X] \right), \\ g^* &= \Pi_{\mathcal{G}} \left(\frac{m}{m + n\lambda} f^* + \frac{n\lambda}{m + n\lambda} \eta \right). \end{aligned}$$

Furthermore, if $\mu \in \mathcal{F}$, $\mu \in \mathcal{G}$, and $\eta \in \mathcal{G}$, then

$$f^* = \mu, \text{ and } g^* = \frac{m}{m + n\lambda} \mu + \frac{n\lambda}{m + n\lambda} \eta.$$

Remark 2.2 (Interpolation Effect). The characterization in Theorem 2.1 highlights how the rich-view model g^* interpolates between the deployment target μ and the rich-view regression function η . In particular, under the realizability assumptions in Theorem 2.1, as $\lambda \rightarrow 0$, the penalty on the rich-view model becomes inactive and the solution satisfies $g^* \rightarrow \mu$, corresponding to a regime in which the privileged information is ignored. On the other hand, as λ increases, g^* places increasing weight on η , recovering behavior akin to a Two-Stage pseudo-labeling procedure. Intermediate values of λ yield an interpolation between these extremes, allowing the method to adapt to the informativeness of the privileged variables.

Correlation Controlled Risk Bound. We now study the statistical behavior of the proposed estimator by analyzing its prediction error. Let $(\hat{f}, \hat{g}) \in \mathcal{F} \times \mathcal{G}$ denote the solution (or an approximate solution) of the empirical problem (2) and let $(f^*, g^*) \in \mathcal{F} \times \mathcal{G}$ denote the minimizer of (3). Our goal is to control the excess risk

$$\mathbb{E}_{\mathcal{D}, X} \left[(\hat{f}(X) - \mu(X))^2 \right],$$

where the expectation is over the training data $\mathcal{D}$ used to construct $\hat{f}$ and an independent test point X .

A key quantity in our analysis is a correlation coefficient that captures the interaction between the estimation errors of the deployment and rich-view models. Specifically, define the estimation errors

$$\hat{e}_f(X) := f^*(X) - \hat{f}(X), \quad \hat{e}_g(Z) := g^*(Z) - \hat{g}(Z).$$

We then define

$$\rho_* = \frac{|\mathbb{E}_{\mathcal{D}, Z} [\hat{e}_f(X) \hat{e}_g(Z)]|}{(\mathbb{E}_{\mathcal{D}, X} [\hat{e}_f^2(X)])^{1/2} (\mathbb{E}_{\mathcal{D}, Z} [\hat{e}_g^2(Z)])^{1/2}} \in [0, 1],$$

where $Z = (X, W)$, which measures the alignment between the residuals of the two estimators, with the convention that $\rho_* = 0$ if either denominator term is zero.

The following corollary, which is a direct consequence of the excess loss decomposition in Theorem C.1, shows that ρ_* plays a central role in controlling the prediction error.

Corollary 2.3 (Correlation Controlled Risk Bound). *If $\mu \in \mathcal{F}$, $\mu \in \mathcal{G}$ (note that functions on $\mathcal{X}$ are implicitly lifted to $\mathcal{Z}$), and $\eta \in \mathcal{G}$, then any estimator $(\hat{f}, \hat{g}) \in \mathcal{F} \times \mathcal{G}$ satisfies*

$$\begin{aligned} \mathbb{E}_{\mathcal{D}, X} \left[(\hat{f}(X) - \mu(X))^2 \right] \\ \leq \frac{\mathbb{E}_{\mathcal{D}} \left[\mathcal{L}(\hat{f}, \hat{g}; \lambda) - \mathcal{L}(f^*, g^*; \lambda) \right]}{\gamma_{n, m, \lambda}(\rho_*)}, \end{aligned}$$

where

$$\gamma_{n, m, \lambda}(\rho_*) := 1 - \frac{m^2}{N(m + n\lambda)} \rho_*^2 \geq \frac{n}{N}.$$

The above bound shows that excess loss in the joint objective translates into prediction error for $\hat{f}$, with a constant that degrades as the correlation ρ_* between the estimation errors increases.

Remark 2.4 (Correlation Score Interpretation). (i) The score ρ_* measures the alignment between the deployment model and rich-view estimation errors; small ρ_* yields a sharper conversion from coupled excess loss to deployment risk and corresponds to privileged information contributing variation not already captured by X , whereas large ρ_* limits the gain from incorporating W . In this regime, leveraging the privileged information can substantially improve the accuracy

of $\hat{f}$. In contrast, when ρ_* is large, the residuals of $\hat{f}$ and $\hat{g}$ are highly correlated, suggesting that W provides little additional information beyond what is already contained in X , and the benefit of incorporating privileged data becomes limited.

(ii) By Cauchy–Schwarz and the law of total variance, it can be shown (see Section C.4 for details) that

$$\rho_*^2 \leq \frac{\mathbb{E}_{\mathcal{D}, X} \left[(\mathbb{E}[\hat{e}_g(Z) | \mathcal{D}, X])^2 \right]}{\mathbb{E}_{\mathcal{D}, Z} [\hat{e}_g^2(Z)]} \leq 1.$$

Furthermore, if $\mu \in \mathcal{F}$, $\mu \in \mathcal{G}$, and $\eta \in \mathcal{G}$, then $\mathbb{E}[g^*(Z) | X] = \mu(X)$, and hence

$$\mathbb{E}[\hat{e}_g(Z) | \mathcal{D}, X] = \mu(X) - \mathbb{E}[\hat{g}(Z) | \mathcal{D}, X].$$

Consequently, ρ_* is small when the conditional rich-view model error $\mathbb{E}_{\mathcal{D}, X} [\mathbb{E}[\hat{e}_g(Z) | \mathcal{D}, X]^2]$ is small relative to the total rich-view model error $\mathbb{E}_{\mathcal{D}, Z} [\hat{e}_g^2(Z)]$. This corresponds to a regime in which the rich-view model $\hat{g}$ leverages privileged information W to capture predictive signal about Y that is not well explained by X alone, thereby reducing the component of the rich-view error that is aligned with X . Hence, ρ_* , and consequently the risk bound, are controlled by a *multiplicative* relative error, in contrast to [Xia & Wainwright \(2024\)](#), where the risk is bounded in terms of an *additive* absolute error $\mathbb{E}_{\mathcal{D}, X} [\mathbb{E}[\hat{e}_g(Z) | \mathcal{D}, X]^2]$. This means that the risk bound degrades more gracefully as $\hat{g}$ worsens, because it depends on the fraction of rich-view error explainable from X rather than on its absolute size.

2.1. Main Risk Bound

Assumption 2.5. The function classes $\mathcal{F}$ and $\mathcal{G}$ are uniformly bounded. Specifically, without loss of generality, there exists a constant $B \geq 1$ such that

$$\sup_{(f, g) \in \mathcal{F} \times \mathcal{G}} \|(f, g)\|_\infty \leq B \quad \text{and} \quad |Y| \leq B, \quad \text{a.s.,}$$

where $\|(f, g)\|_\infty := \max \{\|f\|_\infty, \|g\|_\infty\}$.

Remark 2.6 (On the boundedness assumption). The condition $\sup_{(f, g) \in \mathcal{F} \times \mathcal{G}} \|(f, g)\|_\infty \leq B$ is made for convenience. When $\mathcal{F}$ and $\mathcal{G}$ are unbounded but $|Y| \leq B$ a.s., the same conclusions apply to the truncated estimators $T_B \hat{f}$ and $T_B \hat{g}$, with $T_B(t) := \text{sign}(t) (|t| \wedge B)$ following the truncation argument of [Barron et al. \(2008\)](#). Since $|Y| \leq B$ a.s., truncation can only decrease the (squared) loss, so no rate is lost.

We introduce a complexity term that captures the joint richness of the deployment and rich-view function classes, at scales determined by the sample sizes and the regularization parameter λ . Define

$$\mathfrak{C}_{\mathcal{F} \times \mathcal{G}}^{N, \lambda} := 1 + \sup_{z^N} \log \left(\mathcal{N}_{1, \infty}(\epsilon_N, \mathcal{F} \times \mathcal{G}, z^N) \right),$$

with $\epsilon_N = \frac{1}{1280B(\lambda+1)N}$.

Theorem 2.7 (Risk Bound). *If $\mu \in \mathcal{F}$, $\mu \in \mathcal{G}$, $\eta \in \mathcal{G}$ and Assumption 2.5 holds, then any estimator $(\hat{f}, \hat{g}) \in \mathcal{F} \times \mathcal{G}$ satisfies*

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}, X} \left[(\hat{f}(X) - \mu(X))^2 \right] \\ & \leq \frac{1}{\gamma_{n,m,\lambda}(\rho_*)} \left(\frac{C_1 B^2 (\lambda + 1) \mathfrak{E}_{\mathcal{F} \times \mathcal{G}}^{N,\lambda}}{N} \right. \\ & \quad \left. + 2 \mathbb{E}_{\mathcal{D}} [\hat{\Delta}(\hat{f}, \hat{g})] \right), \end{aligned} \quad (4)$$

where $C_1 > 0$ is a universal constant, and

$$\hat{\Delta}(\hat{f}, \hat{g}) := \hat{\mathcal{L}}(\hat{f}, \hat{g}; \lambda) - \hat{\mathcal{L}}(f^*, g^*; \lambda).$$

Remark 2.8 (Empirical Comparison Term). The term $\hat{\Delta}(\hat{f}, \hat{g})$ is an empirical comparison term, comparing the empirical coupled objective attained by the estimator $(\hat{f}, \hat{g})$ with that of the population reference pair (f^*, g^*) . It is not necessarily nonnegative, since (f^*, g^*) need not minimize the empirical objective. In particular, if $(\hat{f}, \hat{g})$ is an exact empirical minimizer of (2), then

$$\hat{\Delta}(\hat{f}, \hat{g}) \leq 0,$$

so this term can be dropped from (4). For algorithmically constructed estimators, $\hat{\Delta}(\hat{f}, \hat{g})$ records how well the computed pair competes with (f^*, g^*) on the empirical objective; for the alternating forward selection iterates, this comparison is controlled by Theorem 3.1.

Remark 2.9 (Effect of Correlation Score). (i) Note that, for any values of ρ_* and λ , the quantity $\gamma_{n,m,\lambda}(\rho_*)$ satisfies the uniform lower bound $\gamma_{n,m,\lambda}(\rho_*) \geq n/N$. As a consequence,

$$\frac{1}{\gamma_{n,m,\lambda}(\rho_*)} \cdot \frac{1}{N} \leq \frac{1}{n},$$

so the $O(1/N)$ term in (4) is never worse than the $O(1/n)$ dependence obtained from using only the labeled data.

(ii) When the estimation errors of the deployment and rich-view estimators are highly correlated (i.e., ρ_* is close to 1), one has $\gamma_{n,m,\lambda}(\rho_*) \approx 1 - m^2/(N(m+n\lambda))$. In particular, when $\lambda = 0$, this becomes $\gamma_{n,m,0}(\rho_*) \approx n/N$, and hence

$$\frac{1}{\gamma_{n,m,0}(\rho_*)} \cdot \frac{1}{N} \approx \frac{1}{n},$$

so the $O(1/N)$ term in (4) recovers the $O(1/n)$ dependence, obtained when learning solely from n labeled observations.

(iii) More generally, $\gamma_{n,m,\lambda}(\rho_*)$ controls the multiplicative constant in (4) through its dependence on ρ_* . When the deployment and rich-view estimation errors are weakly aligned (small ρ_*), $\gamma_{n,m,\lambda}(\rho_*)$ remains close to one, so the $O(1/N)$ term is not significantly inflated.

(iv) The quantity ρ_* is introduced only for the analysis, not observed from data, and not used to tune the method. Rather,

it summarizes how much of the rich-view error remains explainable from X alone. Constructing a practical proxy for ρ_* that could guide the choice of λ is an interesting direction for future work.

3. High-Dimensional Extension

In Section 1.1, we proposed a coupled training algorithm for jointly estimating the deployment model $\hat{f}$ and the rich-view model $\hat{g}$. In practice, however, directly optimizing over arbitrary functions in infinite-dimensional Hilbert spaces is computationally inefficient, especially when the combined inputs (X, W) are high-dimensional.

To bridge this gap between theory and practice, we adopt a dictionary-based approximation strategy in which functions are represented as sparse linear combinations of elementary functions, or *atoms*, drawn from predefined dictionaries. The resulting procedure is efficient and remains theoretically well-justified. In this section, we introduce a greedy alternating forward selection procedure that incrementally constructs approximations of $(\hat{f}, \hat{g})$ using finite dictionaries. Let $\mathcal{D}_f$ and $\mathcal{D}_g$ be finite, symmetric dictionaries such that $\|\psi\|_{L^2(\hat{P}_N^X)} \leq 1$ for all $\psi \in \mathcal{D}_f$ and $\|\phi\|_{L^2(\hat{P}_N^Z)} \leq 1$ for all $\phi \in \mathcal{D}_g$. For functions $f \in \mathcal{F} = \text{span}(\mathcal{D}_f)$ and $g \in \mathcal{G} = \text{span}(\mathcal{D}_g)$ of the form

$$f = \sum_{\psi \in \mathcal{D}_f} c_\psi \psi, \quad g = \sum_{\phi \in \mathcal{D}_g} c_\phi \phi,$$

we define the associated atomic norms by

$$\begin{aligned} \|f\|_{L^1(\mathcal{D}_f)} &:= \inf \left\{ \sum_{\psi \in \mathcal{D}_f} |c_\psi| : f = \sum_{\psi \in \mathcal{D}_f} c_\psi \psi \right\}, \\ \|g\|_{L^1(\mathcal{D}_g)} &:= \inf \left\{ \sum_{\phi \in \mathcal{D}_g} |c_\phi| : g = \sum_{\phi \in \mathcal{D}_g} c_\phi \phi \right\}. \end{aligned}$$

3.1. Alternating Forward Selection Algorithm

We now describe a greedy algorithm that alternates between updating f and recalibrating g .

The procedure approximately minimizes $\hat{\mathcal{L}}(f, g; \lambda)$ in (2) by alternately updating f and g using forward (stepwise) selection (Hastie et al., 2009, Sec. 3.3.2).

Alternating Forward Selection Algorithm. Initialize $g_0 = 0$. For $k = 1, 2, \dots$, alternate between the following updates.

1. **Update f .** Given the current pseudo-responses $\hat{Y}_j = g_{k-1}(Z_j)$, update f by solving

$$\begin{aligned} (\psi_k, f_k) \in & \underset{\substack{\psi \in \mathcal{D}_f \\ f \in \text{span}(\psi_1, \dots, \psi_{k-1}, \psi)}}{\text{argmin}} \frac{1}{N} \left(\sum_{i=1}^n (Y_i - f(X_i))^2 \right. \\ & \left. + \sum_{j=n+1}^N (g_{k-1}(Z_j) - f(X_j))^2 \right). \end{aligned}$$

2. Recalibrate g . Given the current deployment model f_k , update g by solving

$$(\phi_k, g_k) \in \underset{\substack{\phi \in \mathcal{D}_g \\ g \in \text{span}\{\phi_1, \dots, \phi_{k-1}, \phi\}}}{\text{argmin}} \frac{1}{N} \left(\sum_{j=n+1}^N (g(Z_j) - f_k(X_j))^2 + \lambda \sum_{i=1}^n (Y_i - g(Z_i))^2 \right).$$

An efficient implementation of this procedure is provided in Appendix A (see Algorithm 1). We first establish a convergence rate, then summarize the implementation.

Theorem 3.1 (Alternating Forward Selection Convergence Rate). *For all $k \geq 1$, and all $f \in \text{span}(\mathcal{D}_f)$ and $g \in \text{span}(\mathcal{D}_g)$, the iterates $(\hat{f}, \hat{g}) = (f_k, g_k)$ produced by the Alternating Forward Selection Algorithm satisfy*

$$\hat{\mathcal{L}}(\hat{f}, \hat{g}; \lambda) \leq \hat{\mathcal{L}}(f, g; \lambda) + C_2 \frac{\max\{1, \lambda\} \log(k+1)}{k} \times (\|f\|_{L^1(\mathcal{D}_f)} + \|g\|_{L^1(\mathcal{D}_g)})^2,$$

for a universal constant $C_2 > 0$.

This result shows that the greedy alternating procedure gives sublinear control of the empirical comparison term through its optimization error contribution. After k steps, each component lies in a span generated by at most k selected atoms. Thus, the iteration count records both the achieved optimization accuracy and the complexity of the selected representation along the greedy path.

3.2. Efficient Implementation

We briefly describe an efficient implementation of the Alternating Forward Selection Algorithm; full derivations and pseudocode are deferred to Appendix A. At a high level, each block update (the f -step over $\text{span}(\mathcal{D}_f)$ and the g -step over $\text{span}(\mathcal{D}_g)$) is a least-squares refit over the currently selected atoms. Rather than resolving these refits from scratch, we maintain incremental QR decompositions for the (weighted) design matrices associated with the selected atoms, enabling $O(Nk)$ updates per iteration, where k is the number of selected atoms per block.

Time complexity. Assuming each candidate atom can be evaluated on all N samples in $O(N)$ time, selecting the next atom at iteration k amounts to scanning the remaining candidates in the relevant dictionary and computing their correlations with the current residual after orthogonalization against the selected span (via the maintained QR factors). This yields a total cost up to step k of $O(N(|\mathcal{D}_f| + |\mathcal{D}_g|)k)$. For comparison, a naive version of forward selection over paired atoms (ψ, ϕ) (i.e., scanning a product dictionary) scales as $O(N|\mathcal{D}_f||\mathcal{D}_g|k)$, which is substantially larger. As shown in Theorem 3.1, this computational improvement

comes with only a $\log(k)$ factor difference in the optimization guarantee relative to classical greedy schemes Barron et al. (2008); DeVore & Temlyakov (1996).

4. Experiments

We evaluate the coupled objective on controlled synthetic data and two real-world tasks. The tuning parameter λ interpolates between labeled-only learning on X and the Two-Stage method. Unless otherwise stated, λ is selected by cross-validation using only labeled training data. We compare against (i) supervised learning on $\{(X_i, Y_i)\}_{i=1}^n$ using only X (Baseline), and (ii) the Two-Stage procedure from Section 1. The square-loss synthetic controls match our theory, while the logistic synthetic diagnostic, random-forest regression, and Bank Marketing classification experiments are empirical demonstrations beyond the current proof regime.

Synthetic controls. We first use a linear model with known target $\mu(X)$ and privileged coordinates W containing both signal and nuisance variation. Figure 2 shows that the coupled method improves as the privileged signal strengthens, remains stable as nuisance privileged dimensions increase, and improves as the number of unlabeled samples increases. In contrast, Two-Stage pseudo-labeling can degrade when the rich-view teacher overfits nuisance variation.

Synthetic classification diagnostic. We also evaluate a binary logistic analogue of the coupled objective, replacing squared-loss term in (2) by cross-entropy and using soft rich-view pseudo-labels. As shown in Figure 6 in Appendix B, the test error is minimized at an intermediate value of λ , while the Two-Stage limit over-transfers rich-view noise.

Real-data benchmarks. For regression, we use the UCI Parkinson’s Telemonitoring dataset and predict disease severity from voice measurements. Basic perturbation measures form the deployment view X , while more complex spectral features form the privileged view W . Figure 3 in Appendix B shows the expected U-shaped dependence on λ ; Table 1 reports the same subject-level split with the labeled-only CV choice $\hat{\lambda} = 3$. Due to high variance from subject-level splitting, the main table reports single-seed results for Parkinson’s. The larger Bank task averages 26 seeds. Cross-validation (CV) robustness checks for both are provided in Appendix B.

For classification, we use the Bank Marketing dataset and evaluate by holdout Brier score. Static demographic, contact, and macroeconomic variables form the deployment view X , while current-call and campaign outcome variables form the privileged view W . We evaluate over 26 stratified outer splits, using $n = 200$ labeled examples and $m = 10,000$ unlabeled paired (X, W) examples per split. The tuning parameter λ and all baselines are selected by the same fixed 4-fold cross-validation protocol on the labeled set only. We compare against the X -only Baseline,

Table 1. **Real-data summary.** Lower is better. For Parkinson’s, we report the single-seed subject-level split with λ selected by labeled-only CV. For Bank Marketing, entries are mean holdout Brier scores across 26 outer seeds.

Dataset	Metric	Baseline	Two-Stage	Gen. Distill.	SVM+	Ours
Parkinson’s	MSE	77.34	83.32	—	—	72.60
Bank Marketing	Brier	0.0893	0.0928	0.0907	0.1470	0.0881

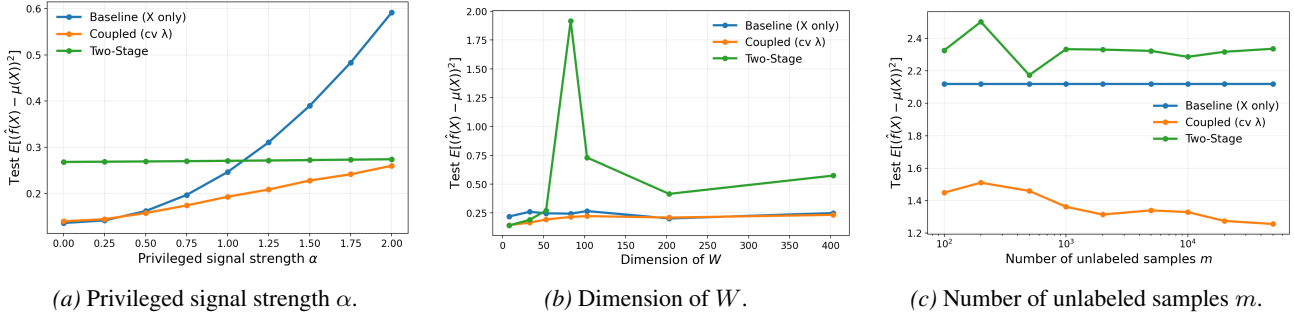


Figure 2. **Synthetic controls.** Test error is $\mathbb{E}[(\hat{f}(X) - \mu(X))^2]$. Coupled training adapts to useful privileged signal, is more stable than Two-Stage under nuisance privileged dimensions, and improves with additional unlabeled data.

Two-Stage pseudo-labeling, a squared-loss generalized distillation baseline, and SVM+ (Vapnik & Vashist, 2009). Coupled Training beats the X -only Baseline and generalized distillation in 22/26 seeds, and beats Two-Stage and SVM+ in 26/26 seeds. Figure 4 in Appendix B shows a representative split, where the cross-validated $\hat{\lambda}$ lands near the interior low-Brier region of the holdout diagnostic curve. Table 1 summarizes the real-data results.

Overall, the experiments support the intended role of Coupled Training, exploiting privileged information when useful while attenuating it when direct pseudo-labeling would transfer nuisance signal into the final X -only predictor. Further implementation details and diagnostics are deferred to Appendix B.

5. Conclusion

We studied learning with privileged features and unlabeled data when the deployment predictor may only use X at test time. The main difficulty in this setting is negative transfer, where a rich-view predictor trained on (X, W) can be useful when W carries a genuine extra signal, but can also mislead the deployment model when the privileged view is weak or noisy. Our coupled objective addresses this by jointly fitting the deployment model and the rich-view model, rather than treating the rich-view predictor as a fixed teacher.

Our estimator interpolates between labeled-only learning (weak coupling) and Two-Stage pseudo-labeling (strong coupling). This interpolation is reflected both in the population characterization and in the empirical U-shaped dependence on λ observed across our synthetic and real data experiments. In addition, for the high-dimensional setting, we developed a dictionary-based AFS procedure and established a sublinear optimization guarantee, which signifi-

cantly reduces the time complexity.

Limitations. Our theory covers square-loss regression with convex function classes, so the random-forest and classification experiments are empirical evidence beyond the present proof setting. The method also assumes paired unlabeled samples with privileged features; extensions to misspecification, distribution shift, semi-paired data, and broader classification guarantees remain future work.

Acknowledgements

We extend our special thanks to the anonymous reviewers for their insightful comments and suggestions that greatly enhanced this work. Hagrass gratefully acknowledges financial support from the Schmidt DataX Fund at Princeton University made possible through a major gift from the Schmidt Futures Foundation. Klusowski is grateful for support from the National Science Foundation through NSF CAREER DMS-2239448 and the Alfred P. Sloan Foundation through a Sloan Research Fellowship.

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 of which we feel must be specifically highlighted here.

References

- Agarwal, A., Anandkumar, A., Jain, P., Netrapalli, P., and Tandon, R. Learning Sparsely Used Overcomplete Dictionaries. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pp. 123–137. PMLR, 2014.
- Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A.

- Approximation and Learning by Greedy Algorithms. *The Annals of Statistics*, 36(1):64 – 94, 2008.
- Chakraborty, A. and Cai, T. Efficient and Adaptive Linear Regression in Semi-Supervised Settings. *The Annals of Statistics*, 46(4):1541–1572, 2018.
- Chapelle, O., Schölkopf, B., and Zien, A. (eds.). *Semi-Supervised Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2006. ISBN 978-0-262-03358-9.
- Chatterji, N. S. and Bartlett, P. L. Alternating minimization for dictionary learning: Local Convergence Guarantees. In *Conference on Learning Theory*, 2017.
- DeVore, R. A. and Temlyakov, V. N. Some remarks on greedy algorithms. *Advances in Computational Mathematics*, 5(1):173–187, 1996. doi: 10.1007/BF02124742.
- Ding, D. Y., Li, S., Narasimhan, B., and Tibshirani, R. Co-operative learning for multiview analysis. *Proceedings of the National Academy of Sciences*, 119(38):e2202113119, 2022.
- Fries, J. A., Varma, P., Chen, V. S., et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac MRI sequences. *Nature Communications*, 10(1):3111, 2019.
- Grippo, L. and Sciandrone, M. On the convergence of the block nonlinear Gauss–Seidel method under convex constraints. *Operations Research Letters*, 26(3):127–136, 2000. ISSN 0167-6377. doi: 10.1016/S0167-6377(99)00074-7.
- Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.
- Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.
- Hogan, J. W., Roy, J., and Korkontzelou, C. Handling dropout in longitudinal studies. *Statistics in Medicine*, 23(9): 1455–1497, 2004. doi: 10.1002/sim.1728.
- Hou, J., Guo, Z., and Cai, T. Surrogate Assisted Semi-supervised Inference for High Dimensional Risk Prediction. *Journal of Machine Learning Research*, 24(265): 1–58, 2023.
- Joachims, T. Transductive Inference for Text Classification Using Support Vector Machines. In *Proceedings of the 16th International Conference on Machine Learning*, volume 99, pp. 200–209, 1999.
- Koh, P. W., Sagawa, S., Marklund, H., et al. WILDS: A Benchmark of In-the-Wild Distribution Shifts. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 5637–5664. PMLR, 2021.
- Laine, S. and Aila, T. Temporal Ensembling for Semi-Supervised Learning. In *International Conference on Learning Representations*, 2017.
- Lee, D.-H. Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. In *Workshop: Challenges in Representation Learning*, volume 3, pp. 896, 2013.
- Natarajan, N., Dhillon, I. S., Ravikumar, P. K., and Tewari, A. Learning with Noisy Labels. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1196–1204, 2013.
- Rajkomar, A., Dean, J., and Kohane, I. Machine Learning in Medicine. *New England Journal of Medicine*, 380(14): 1347–1358, 2019.
- Ratner, A., Bach, S., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. In *Proceedings of the VLDB Endowment*, volume 11, pp. 269, 2017.
- Ratner, A. J. et al. Data Programming: Creating Large Training Sets, Quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016.
- Robinson, J., Jegelka, S., and Sra, S. Strength from Weakness: Fast Learning Using Weak Supervision. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 8127–8136, 2020.
- Ruetz, S. and Schnass, K. Convergence of alternating minimisation algorithms for dictionary learning. *arXiv preprint arXiv:2304.01768*, 2023.
- Sindhwani, V., Niyogi, P., and Belkin, M. A Co-Regularization Approach to Semi-Supervised Learning with Multiple Views. In *Proceedings of the Workshop on Learning with Multiple Views*, 22nd ICML, 2005.
- Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. Learning from Noisy Labels with Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2022.
- Vapnik, V. and Vashist, A. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. ISSN 0893-6080. doi: <https://doi.org/10.1016/j.neunet.2009.06.042>. Advances in Neural Networks Research: IJCNN2009.

- Vapnik, V., Izmailov, R., et al. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16(1):2023–2049, 2015.
- Wang, K. Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift. *arXiv preprint arXiv:2302.10160*, 2023.
- Xia, E. and Wainwright, M. J. Prediction Aided by Surrogate Training. *arXiv preprint arXiv:2412.09364*, 2024.
- Zhang, A., Brown, L. D., and Cai, T. T. Semi-Supervised Inference: General Theory and Estimation of Means. *The Annals of Statistics*, 47(5):2538–2566, 2019.
- Zhang, Y. and Bradic, J. High-dimensional semi-supervised learning: in search of optimal inference of the mean. *Biometrika*, 109(2):387–403, 2021.
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

A. Efficient Implementation of Alternating Forward Selection

Recall from Section 1 that we observe n labeled and m unlabeled samples with $N = n + m$, and that $\mathcal{Z} = \mathcal{X} \times \mathcal{W}$. Let $\pi_X : \mathcal{Z} \rightarrow \mathcal{X}$ be the canonical projection $\pi_X(x, w) = x$. Define the finite measures on $\mathcal{Z}$ by

$$\tilde{P}^{(1)} := \sum_{i=1}^n \delta_{Z_i}, \quad \tilde{P}^{(2)} := \sum_{j=n+1}^N \delta_{Z_j}, \quad \tilde{P}^{(3)} := \lambda \tilde{P}^{(1)},$$

and normalize by N .

$$\hat{P}_N^{(\ell)} := \frac{1}{N} \tilde{P}^{(\ell)}, \quad \ell \in \{1, 2, 3\}.$$

Note that $\hat{P}_N^{(1)} + \hat{P}_N^{(2)} = \hat{P}_N^Z$. We will work exclusively with Z -indexed norms by viewing the deployment class as a subspace of functions on $\mathcal{Z}$. Define the lifted deployment space

$$\mathcal{H}_f := \{ f \in L^2(\hat{P}_N^Z) : f \text{ is } \sigma(\pi_X)\text{-measurable} \},$$

i.e., $f \in \mathcal{H}_f$ implies $f(x, w)$ depends only on x (equivalently, there exists $\bar{f} : \mathcal{X} \rightarrow \mathbb{R}$ such that $f(x, w) = \bar{f}(x)$ $\hat{P}_N^Z$ -a.e.).

Let $Q_g := \hat{P}_N^{(2)} + \hat{P}_N^{(3)}$ and set

$$\mathcal{H}_g := L^2(Q_g).$$

We assume the function classes live in these Hilbert spaces, $\mathcal{F} \subset \mathcal{H}_f, \mathcal{G} \subset \mathcal{H}_g$. To keep notation light, we identify the model class $\mathcal{F}$ with its lift to $\mathcal{Z}$, namely, we view each $f \in \mathcal{F}$ as a function $f(x, w)$ that is $\sigma(\pi_X)$ -measurable (hence depends only on x , i.e., is constant in w given x). We use $\mathcal{D}_f$ in the same spirit, and throughout the appendix it refers to a dictionary in the lifted space $\mathcal{H}_f$. With this convention, for $f \in \mathcal{H}_f, g \in \mathcal{H}_g$, the empirical objective can be written as

$$\hat{\mathcal{L}}(f, g; \lambda) := \|Y - f\|_{L^2(\hat{P}_N^{(1)})}^2 + \|f - g\|_{L^2(\hat{P}_N^{(2)})}^2 + \|Y - g\|_{L^2(\hat{P}_N^{(3)})}^2.$$

Let

$$\mathcal{K}_1 := L^2(\hat{P}_N^{(1)} + \hat{P}_N^{(2)}), \quad \mathcal{K}_2 := L^2(\hat{P}_N^{(2)} + \hat{P}_N^{(3)}), \quad \mathcal{K} := \mathcal{K}_1 \times \mathcal{K}_2.$$

For $u = (u_1, u_2)$ and $v = (v_1, v_2)$ in $\mathcal{K}$, define

$$\langle u, v \rangle_\star = \langle u_1, v_1 \rangle_{L^2(\hat{P}_N^{(1)})} + \langle u_1 - u_2, v_1 - v_2 \rangle_{L^2(\hat{P}_N^{(2)})} + \langle u_2, v_2 \rangle_{L^2(\hat{P}_N^{(3)})}.$$

This bilinear form is positive semidefinite. Let

$$\mathcal{N}_\star := \{u \in \mathcal{K} : \|u\|_\star = 0\}$$

and define the quotient Hilbert space

$$\bar{\mathcal{K}} := \mathcal{K} / \mathcal{N}_\star.$$

We write $[u]$ for the equivalence class of u , although, when no confusion is possible, we suppress brackets and use the same notation for a representative and its quotient class. The deployment and rich-view model spaces are embedded as

$$\mathcal{H}^\star := \{[(f, g)] : f \in \mathcal{H}_f, g \in \mathcal{H}_g\} \subset \bar{\mathcal{K}}.$$

Below, all projections, residuals, spans, and orthogonality statements are taken in $\bar{\mathcal{K}}$ with respect to $\langle \cdot, \cdot \rangle_\star$.

Let Y be any extension of the labeled responses to all N sample points, with $Y(Z_i) = Y_i$ for $i \leq n$. Define

$$T := (Y, Y) \in \bar{\mathcal{K}}.$$

The choice of Y on unlabeled samples is irrelevant, since the unlabeled term depends on $(Y - f) - (Y - g) = g - f$. Hence, for every f, g ,

$$\|T - (f, g)\|_\star^2 = \|Y - f\|_{L^2(\hat{P}_N^{(1)})}^2 + \|f - g\|_{L^2(\hat{P}_N^{(2)})}^2 + \|Y - g\|_{L^2(\hat{P}_N^{(3)})}^2 = \hat{\mathcal{L}}(f, g; \lambda).$$

Without loss of generality, we suppose that the dictionaries $\mathcal{D}_f$ and $\mathcal{D}_g$ are bounded by 1 in the *pooled* empirical norm, i.e.,

$$\sup_{\psi \in \mathcal{D}_f} \|\psi\|_{L^2(\hat{P}_N^Z)} \leq 1, \quad \sup_{\phi \in \mathcal{D}_g} \|\phi\|_{L^2(\hat{P}_N^Z)} \leq 1.$$

Algorithm 1 Alternating Forward Selection

Input. Dictionaries $\mathcal{D}_f^*, \mathcal{D}_g^*$; initial values $(f_0, g_0) = (0, 0)$, $S_0^f = S_0^g = \{0\}$; target (Y, Y) .

for $k = 1, 2, \dots$ **do**

 Compute the residual.

$$r_k \leftarrow (Y, Y) - (f_{k-1}, g_{k-1}).$$

Step 1. f-selection and projection

 Select

$$\psi_k^* = (\psi_k, 0) \in \underset{\substack{a \in \mathcal{D}_f^* \\ \|\Pi_{S_{k-1}^f}^\perp a\|_* > 0}}{\operatorname{argmax}} \frac{\langle r_k, \Pi_{S_{k-1}^f}^\perp a \rangle_*}{\|\Pi_{S_{k-1}^f}^\perp a\|_*}.$$

 Update

$$\begin{aligned} S_k^f &\leftarrow \operatorname{span}(\psi_1^*, \dots, \psi_k^*), \\ (f_k, g_{k-1}) &\leftarrow (f_{k-1}, g_{k-1}) + \Pi_{S_k^f} r_k, \end{aligned}$$

 and set

$$r_k^{(g)} \leftarrow r_k - \Pi_{S_k^f} r_k.$$

Step 2. g-selection and projection

 Select

$$\phi_k^* = (0, \phi_k) \in \underset{\substack{b \in \mathcal{D}_g^* \\ \|\Pi_{S_{k-1}^g}^\perp b\|_* > 0}}{\operatorname{argmax}} \frac{\langle r_k^{(g)}, \Pi_{S_{k-1}^g}^\perp b \rangle_*}{\|\Pi_{S_{k-1}^g}^\perp b\|_*}.$$

 Update

$$\begin{aligned} S_k^g &\leftarrow \operatorname{span}(\phi_1^*, \dots, \phi_k^*), \\ (f_k, g_k) &\leftarrow (f_k, g_{k-1}) + \Pi_{S_k^g} r_k^{(g)}, \end{aligned}$$

end for

In the following, we use canonical embeddings to construct the dictionary for the Hilbert space $\mathcal{H}^*$ based on $\mathcal{D}_f$ and $\mathcal{D}_g$. Define embedded versions.

$$\mathcal{D}_f^* := \{(\psi, 0) : \psi \in \mathcal{D}_f\}, \quad \mathcal{D}_g^* := \{(0, \phi) : \phi \in \mathcal{D}_g\}.$$

Assume the union $\mathcal{D}^* := \mathcal{D}_f^* \cup \mathcal{D}_g^*$ is balanced, i.e., whenever $\zeta \in \mathcal{D}^*$, we have $-\zeta \in \mathcal{D}^*$. Since $\hat{P}_N^{(1)} + \hat{P}_N^{(2)} = \hat{P}_N^Z$, we have

$$\|(\psi, 0)\|_*^2 = \|\psi\|_{L^2(\hat{P}_N^{(1)})}^2 + \|\psi\|_{L^2(\hat{P}_N^{(2)})}^2 = \|\psi\|_{L^2(\hat{P}_N^Z)}^2 \leq 1.$$

Moreover,

$$\|(0, \phi)\|_*^2 = \|\phi\|_{L^2(\hat{P}_N^{(2)})}^2 + \|\phi\|_{L^2(\hat{P}_N^{(3)})}^2 = \frac{1}{N} \sum_{j=n+1}^N \phi(Z_j)^2 + \frac{\lambda}{N} \sum_{i=1}^n \phi(Z_i)^2 \leq \max\{1, \lambda\} \|\phi\|_{L^2(\hat{P}_N^Z)}^2 \leq \max\{1, \lambda\}.$$

Denote $R_\lambda := \max\{1, \sqrt{\lambda}\}$. Thus every embedded atom satisfies $\|a\|_* \leq R_\lambda$. Throughout this subsection, Π_S and $\Pi_S^\perp$ denote the orthogonal projections onto a subspace $S \subset \bar{\mathcal{K}}$ and its orthogonal complement, respectively, with respect to the inner product $\langle \cdot, \cdot \rangle_*$.

B. Implementation Details and Additional Experiments

B.1. Parkinson's (Telemonitoring)

Dataset. We use the UCI Parkinson's Telemonitoring dataset to evaluate Coupled Training. The target is `total_UPDRS`, which typically requires in-person clinical assessment, so labeled data are limited. Each row is a single voice recording, and each `subject` may contribute multiple recordings. Since recording conditions at deployment are less controlled, we treat some acoustic descriptors as privileged features available only during training.

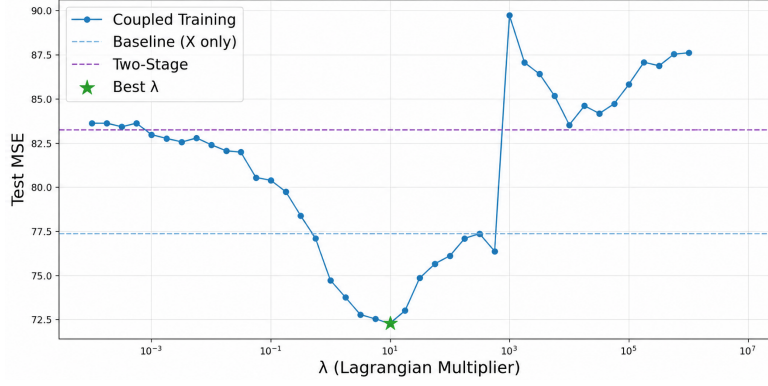


Figure 3. **Parkinson’s dataset.** Test MSE versus λ .

Feature split. We partition covariates into deployment features X and privileged features W to model the fact that some high-fidelity acoustic descriptors are only reliable in controlled clinical conditions.

- **Deployment features X (10)** are {age, sex, test_time, Jitter(%), Jitter(Abs), Jitter:RAP, Jitter:PPQ5, Jitter:DDP, Shimmer, Shimmer(dB)}.
- **Privileged features W (9)** are {Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA, NHR, HNR, RPDE, DFA, PPE}.

Here W contains higher-fidelity acoustic descriptors that are reliable primarily in controlled clinical recordings, whereas X contains features assumed available at deployment. We use W only during training to improve the deployment predictor $\hat{f}(X)$.

Data split. To avoid leakage across repeated recordings from the same subject, we split the data at the subject level. We hold out 10% of subjects as a test set (`seed=202`). From the remaining subjects, we label 15% and treat the rest as unlabeled.

Preprocessing. All preprocessing is fit on the training set only. We standardize non-demographic voice features and apply PCA with 12 components (or fewer if limited by dimension), standardize demographic features (age and test time), and leave sex unchanged. We also standardize Y using statistics computed from labeled training targets.

Models and hyperparameters. We use a random forest regressor with 1400 trees and maximum depth 12 (minimum leaf size 4); remaining settings are kept fixed across methods.

Training procedure. Baseline trains f on labeled data only using $X \rightarrow Y$. Two-Stage trains a teacher g on labeled data using $(X, W) \rightarrow Y$, pseudo-labels unlabeled data with $\tilde{Y}_U = g(X_U, W_U)$, then trains a student on $(X_L \cup X_U) \rightarrow (Y_L \cup \tilde{Y}_U)$.

Coupled Training runs the coupled training updates for up to 15 iterations. For each λ in a log-spaced grid $\{10^{-4}, \dots, 10^6\}$ with 41 points, we initialize f from the labeled baseline and initialize g by fitting to a constant target $g_0 \equiv 0$. In this random-forest experiment we use the penalized Lagrangian form of the square-loss objective,

$$\sum_{i \in L} (Y_i - f(X_i))^2 + \sum_{j \in U} (g(X_j, W_j) - f(X_j))^2 + \lambda \sum_{i \in L} (Y_i - g(X_i, W_i))^2,$$

with all targets standardized using the labeled training responses. Given the current g , the f -update is a random-forest regression of the targets $(Y_L, g(X_U, W_U))$ on the deployment features (X_L, X_U) . Given the current f , the g -update is a weighted random-forest regression of the targets $(Y_L, f(X_U))$ on the rich-view features $((X_L, W_L), (X_U, W_U))$, using weight λ for labeled samples and weight 1 for unlabeled samples. We employ early stopping based on stabilization of the disagreement on unlabeled data (patience 2). We then evaluate the resulting $\hat{f}(X)$ on the held-out test set.

Evaluation. All models are trained to predict scaled Y ; at evaluation, we inversely transform predictions back to the original UPDRS scale, clip to $[0, 100]$, and report test mean squared error (MSE).

Figure 3 reports test MSE as a function of λ ; performance is best at an intermediate value (around 10^1).

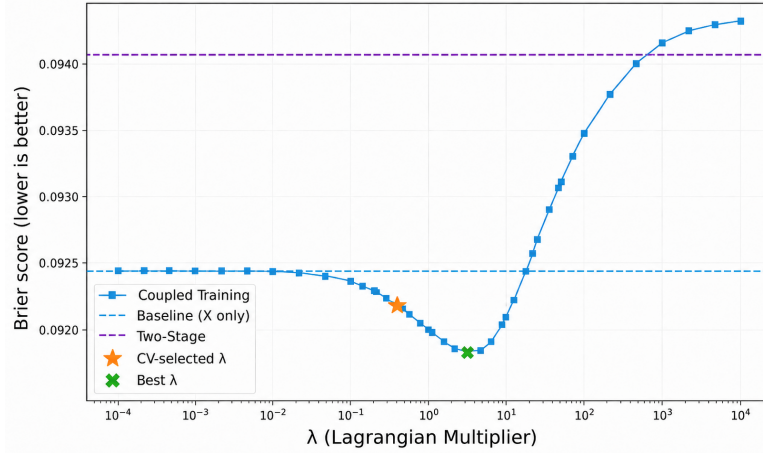


Figure 4. **Bank Marketing dataset.** Holdout Brier score versus λ .

Cross-validation for λ . We select $\hat{\lambda}$ using 5-fold cross-validation on the labeled set only, splitting by subject to prevent leakage (GROUPKFOLD). In each fold, we train Coupled Training using the fixed unlabeled pool $\mathcal{D}_U$ and evaluate the validation MSE of the deployment model $\hat{f}_{\lambda}$ on the held-out labeled fold. We take $\hat{\lambda} \in \arg \min_{\lambda \in \Lambda} \text{MSE}_{\text{val}}(\hat{f}_{\lambda})$, where $\Lambda = \{10^{-4}, \dots, 10^6\}$ is a log-spaced grid with 41 points and $(\hat{f}_{\lambda}, \hat{g}_{\lambda})$ denotes the models returned by Coupled Training when run with tuning parameter λ . The main-table result uses this CV-selected $\hat{\lambda}$; because the dataset has few independent subjects after subject-level splitting, we treat this as the primary diagnostic split rather than averaging a high-variance mean. As a robustness check over seeds $\{202, 42, 123, 999, 777, 2023, 2024, 888\}$, CV-selected Coupled Training beats both Baseline and Two-Stage in 6/8 splits, with wins at seeds $\{202, 999, 777, 2023, 2024, 888\}$.

B.2. Bank Marketing

Dataset. We evaluate Coupled Training on the UCI Bank Marketing dataset, a binary classification task where the goal is to predict whether a client subscribes to a term deposit. The dataset contains 41,188 examples and has a positive class rate of approximately 0.113. We report Brier score, so lower values are better.

Feature split. We use a train-time privileged information split motivated by the distinction between pre-deployment information and information that is tied to the current campaign call or campaign outcomes.

- **Deployment features X** are `{age, job, marital, education, default, housing, loan, contact, month, day_of_week, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed}`.
- **Privileged features W** are `{duration, pdays, previous, poutcome}`.

The final deployed predictor uses only X . The privileged view W is used only during training.

Data split. For each outer seed in

$$\{2034, 2035, \dots, 2059\},$$

we use a stratified random 80/20 train–test split. From the training pool, we sample $n = 200$ labeled examples stratified by class and $m = 10,000$ unlabeled examples from the remaining training examples. In each split, the labeled set contains 23 positives and 177 negatives.

Preprocessing. All preprocessing is fitted on the training pool only. Numeric features are standardized, categorical features are one-hot encoded with unknown categories ignored at test time, and the transformed design matrices are dense. The test set is never used for preprocessing, hyperparameter selection, or model fitting.

Training-time corruption of privileged features. To model noisy or imperfect train-time privileged information, we corrupt W only on training rows. The fixed corruption preset is `mix_plus_noise`, where a fraction 0.35 of the training rows has its privileged feature block row-mixed, then featurewise dropout with probability 0.10 is applied, followed by Gaussian noise with standard deviation 0.15. The same corruption protocol is used for Coupled Training, Two-Stage, and squared-loss generalized distillation whenever those methods use W .

Coupled model and primary baselines. The X -only Baseline, Two-Stage, squared-loss generalized distillation, and Coupled Training models all use linear squared-loss predictors with an unpenalized intercept. The fixed ridge parameters for Coupled Training and Two-Stage are $\alpha_f = 300$ for the deployment model and $\alpha_g = 0.01$ for the rich-view model. The agreement weight on unlabeled examples is 1, and the Two-Stage pseudo-label weight is also 1.

Let $\bar{X}$ denote X augmented with an intercept and let $\bar{Z}$ denote (X, W) augmented with an intercept. For each λ , the coupled model solves

$$\min_{\beta, \gamma} \sum_{i \in L} (Y_i - \bar{X}_i^\top \beta)^2 + \sum_{j \in U} (\bar{X}_j^\top \beta - \bar{Z}_j^\top \gamma)^2 + \lambda \sum_{i \in L} (Y_i - \bar{Z}_i^\top \gamma)^2 + \alpha_f \|\beta_{-0}\|_2^2 + \alpha_g \|\gamma_{-0}\|_2^2,$$

where the subscript -0 excludes the intercept.

Cross-validation for λ . We select λ by fixed 4-fold stratified cross-validation on the 200 labeled examples only. The unlabeled examples are used for training inside each fold, but their labels are never used for selection. The grid is

$$\Lambda = \text{unique} \left(\text{logspace}(-4, 4, 25) \cup \text{logspace}(-1, 2, 21) \right).$$

After selecting $\hat{\lambda}$, we refit on the full labeled set and the same unlabeled pool, then evaluate the X -only deployment predictor on the held-out test set.

Squared-loss generalized distillation. We include a same-loss generalized-distillation baseline. The teacher is a ridge-regularized linear predictor trained on a privileged train-time view Z , where Z is selected from either W alone or $[X, W]$. For a teacher ridge parameter α_T , the teacher solves

$$\hat{\gamma} \in \arg \min_{\gamma} \sum_{i \in L} (Y_i - \bar{Z}_i^\top \gamma)^2 + \alpha_T \|\gamma_{-0}\|_2^2.$$

It produces soft teacher targets

$$q_i = \bar{Z}_i^\top \hat{\gamma}.$$

The student is an X -only ridge model trained by soft squared-loss distillation.

$$\min_{\beta} \sum_{i \in L} (Y_i - \bar{X}_i^\top \beta)^2 + a_L \sum_{i \in L} (q_i - \bar{X}_i^\top \beta)^2 + a_U \sum_{j \in U} (q_j - \bar{X}_j^\top \beta)^2 + \alpha_S \|\beta_{-0}\|_2^2.$$

There is no hard pseudo-label term. We select the teacher view, teacher ridge parameter, student ridge parameter, and soft-distillation weights by the same fixed 4-fold labeled-only CV protocol used for the coupled model. The narrowed search grid is

$$Z \in \{W, [X, W]\}, \quad \alpha_T \in \{0.003, 0.01, 0.03\}, \quad \alpha_S \in \{90, 300, 900\},$$

and

$$(a_L, a_U) \in \{(0, 0.5), (0, 1), (0.25, 0.5), (0.25, 1), (0.5, 0.5), (0.5, 1)\}.$$

This grid contains the ordinary squared-loss Two-Stage baseline as the special case

$$Z = [X, W], \quad \alpha_T = \alpha_g = 0.01, \quad \alpha_S = \alpha_f = 300, \quad (a_L, a_U) = (0, 1).$$

Vapnik SVM+. We also compare against Vapnik SVM+, a labeled-only privileged information baseline. SVM+ uses W only for the labeled training examples and predicts with X only at test time. Since classical SVM+ is not a semi-supervised method, it does not use the unlabeled pool for learning. We tune

$$C \in \{0.1, 1, 10\}, \quad \gamma \in \{0.1, 1, 10\}$$

by the same fixed 4-fold labeled-only CV protocol, then refit on all 200 labeled examples. We use the classical SVM+ limit $\Delta = \infty$ and Platt calibration with calibration parameter $C = 1$ to obtain probabilities for Brier-score evaluation.

Evaluation. For all methods, the final test-time predictor is an X -only predictor. We clip predicted probabilities to $[10^{-6}, 1 - 10^{-6}]$ and report holdout Brier score. All model choices are made without using holdout labels.

Table 2. Bank Marketing detailed comparison. Means and confidence intervals are computed across 26 outer seeds. “Coupled gain” is the mean Brier-score improvement of the CV-selected coupled model over the corresponding method; positive values favor Coupled Training.

Method	Uses unlabeled?	Mean Brier	95% CI	Coupled gain
Baseline (X only)	No	0.0893	[0.0888, 0.0897]	0.0011
Two-Stage	Yes	0.0928	[0.0918, 0.0939]	0.0047
Gen. Distill. (squared)	Yes	0.0907	[0.0894, 0.0921]	0.0026
Vapnik SVM+	No	0.1470	[0.1221, 0.1720]	0.0589
Coupled, CV-selected λ	Yes	0.0881	[0.0874, 0.0889]	–

Results. Table 2 reports the 26-seed comparison. The CV-selected coupled model obtains mean holdout Brier score 0.0881, improving over the X -only Baseline, Two-Stage, squared-loss generalized distillation, and Vapnik SVM+. Squared-loss generalized distillation improves over Two-Stage, with mean Brier score 0.0907 compared with 0.0928, but remains worse than Coupled Training. The mean Brier-score gain of Coupled Training is 0.0011 against the Baseline, 0.0047 against Two-Stage, 0.0026 against squared-loss generalized distillation, and 0.0589 against Vapnik SVM+. Coupled Training beats the Baseline in 22/26 seeds, Two-Stage in 26/26 seeds, squared-loss generalized distillation in 22/26 seeds, and Vapnik SVM+ in 26/26 seeds. The mean regret of the CV-selected $\hat{\lambda}$ relative to the per-seed holdout-best λ is 0.00125 Brier score, with median regret 0.00050.

B.3. PneumoniaMNIST Experiment with AFS-Based Dictionary Selection

Dataset. We use Algorithm 1 as a dictionary-selection stage on PneumoniaMNIST, a binary classification task with 28×28 grayscale images. We use the flattened image as deployment features $X \in \mathbb{R}^{784}$ and construct privileged features $W \in \mathbb{R}^{18}$ available only during training, with the first coordinate of W a noisy version of Y and the remaining ones random linear projections of X . The label $Y \in \{0, 1\}$ indicates pneumonia.

Feature split. Let $X \in \mathbb{R}^{N \times 784}$ be pixel features scaled to $[0, 1]$. We generate $W \in \mathbb{R}^{N \times 18}$ by (i) $W_{:,1} = Y + \varepsilon$ with i.i.d. $\varepsilon \sim \mathcal{N}(0, 0.3^2)$, and (ii) $W_{:,2:18} = XP$ where $P \in \mathbb{R}^{784 \times 17}$ has i.i.d. entries $\mathcal{N}(0, 0.1^2)$ (fixed given the seed).

Data split. We use a single stratified split with $n = 100$ labeled points, $m = 2000$ unlabeled points, and $n_{\text{test}} = 1500$ test points (seed=42).

Preprocessing. For all linear/kernel methods, we standardize both X and W using `StandardScaler` fit on labeled+unlabeled data only (the test set is never used for fitting). We also feed standardized W to the separately trained rich-view teachers described below.

AFS selection stage and dictionaries. For each of 10 log-spaced values of λ from 10^{-2} to 10^2 , we run Algorithm 1 in $\mathcal{H}^* = \mathcal{H}_f \times \mathcal{H}_g$ (Section 3.2) for $K = 70$ iterations with squared loss. The internal iterates (f_k, g_k) drive the successive atom selections. After iteration K , we retain only the selected deployment span S_K^f for the post-selection refit below; neither f_K nor g_K is used as the final deployment predictor or as the refit teacher. We consider the following dictionary constructions.

- **Linear AFS (random-feature dictionary).** Embedded dictionaries use fixed Gaussian random projections with sizes `dict_size_f` = `dict_size_g` = 2048 for f and g respectively.
- **Kernelized AFS (RBF kernel dictionary).** Centers are all labeled points plus up to 500 randomly chosen unlabeled points. Kernel bandwidths γ_f (for X) and γ_g (for (X, W)) are initialized by the median heuristic (computed on up to 600 samples), and tuned by stratified 5-fold CV on labeled data, maximizing AUROC over $\gamma \in \gamma_0 \cdot \{0.25, 0.5, 1, 2, 4\}$ and $\alpha \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$.

Post-selection teacher and refit. Separately from the selection-stage iterate g_K , we fit a labeled-only rich-view teacher g_T using $(X_L, W_L) \rightarrow Y_L$ (ridge in the linear case; kernel ridge in the kernel case with tuned α_g). We form pseudo-targets $\tilde{Y}_U = g_T(X_U, W_U)$ and fit the final deployment predictor $\hat{f}_{\text{refit}}$ by ridge regression with $\alpha_{\text{refit}} = 10^{-3}$ on $(X_L \cup X_U) \rightarrow (Y_L \cup \tilde{Y}_U)$, restricted to S_K^f . Thus, Algorithm 1 determines the deployment dictionary span, whereas this separate ridge refit determines the reported coefficients. Theorem 3.1 concerns the selection-stage iterates, not the post-selection refit or its test metrics.

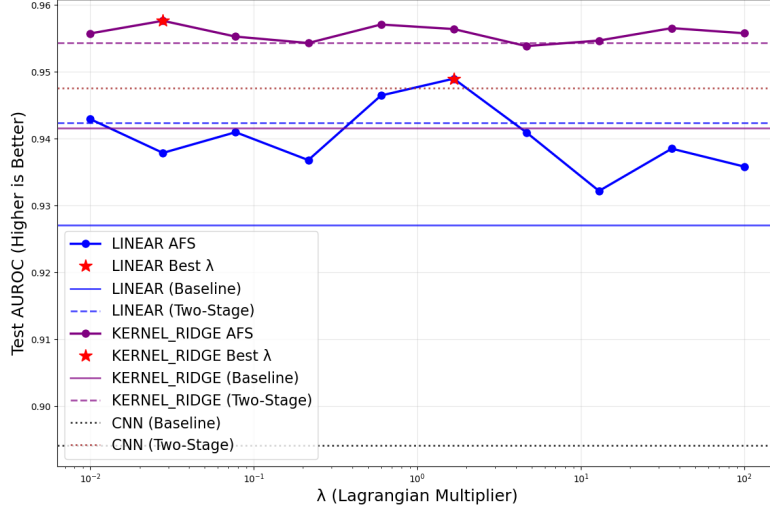


Figure 5. **PneumoniaMNIST**. Test AUROC of the post-selection deployment refit versus λ . Algorithm 1 is used only to select the deployment dictionary span.

Training procedure. Baseline trains f on labeled data only using $X \rightarrow Y$ (ridge for linear; kernel ridge with tuned (γ_f, α_f) for kernel). Two-Stage uses the same labeled-only rich-view teacher construction to pseudo-label the unlabeled data, but fits its deployment student in the full model class rather than restricting it to the AFS-selected span S_K^f . For reference, we also report a CNN baseline (image-only student) and a CNN Two-Stage teacher–student pipeline where the teacher fuses image features with standardized W , pseudo-labels unlabeled data (threshold 0.5), and retrain the image-only student; we use Adam with learning rate 10^{-3} , batch size 32, and 30 epochs.

Evaluation. For the AFS-based methods, the reported deployment score is $\hat{f}_{\text{refit}}(X)$. Models output real-valued scores; we apply a sigmoid to obtain probabilities and report test AUROC (primary), along with accuracy at threshold 0.5 and probability MSE against $\{0, 1\}$ targets.

Results. Figure 5 shows the test AUROC of the post-selection refits as a function of λ . Coupled Training with AFS-based selection performs better than the Baseline and Two-Stage methods for many λ values, with the strongest results at intermediate λ . The effect is especially clear for linear models. For a subset of λ choices, linear AFS-based selection even outperforms both CNN Baseline and Two-Stage.

B.4. Controlled Synthetic Linear Experiments

Dataset. We generate controlled synthetic data to isolate three effects that are central to the behavior of Coupled Training, namely the strength of the privileged signal, the amount of nuisance variation in the privileged view, and the number of unlabeled samples. The data-generating process separates the deployment signal from the privileged signal. Specifically, we sample

$$X \sim \mathcal{N}(0, I_{d_X}), \quad H \sim \mathcal{N}(0, I_q), \quad V \sim \mathcal{N}(0, I_{d_{\text{noise}}}),$$

where H contains the informative latent privileged factors and V contains nuisance privileged coordinates. We then form

$$W_{\text{sig}} = \rho_{XW} X A + \sqrt{1 - \rho_{XW}^2} H, \quad W = (W_{\text{sig}}, V),$$

where the columns of A are normalized. The response is generated as

$$Y = X^\top \beta + \alpha H^\top \theta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

Here α controls the strength of the privileged signal. Since H is not available at deployment, the deployment target is

$$\mu(X) = X^\top \beta.$$

Thus, unlike ordinary test MSE against noisy labels, this construction lets us evaluate the theory-aligned estimation error of the learned deployment predictor.

Feature split. The deployment view is $X \in \mathbb{R}^{d_X}$ and the privileged view is $W \in \mathbb{R}^{q+d_{\text{noise}}}$. The first q coordinates of W are correlated with the latent privileged signal, while the remaining d_{noise} coordinates are nuisance dimensions. This design allows us to test whether the rich-view model helps when the privileged signal is informative, and whether it becomes harmful when the privileged view contains many irrelevant coordinates.

Data split. For each random seed, we independently sample labeled, unlabeled, and test sets from the same ground-truth model. Unless otherwise stated, we use $d_X = 10$, $q = 3$, $\rho_{XW} = 0.7$, and $\sigma = 1$. The signal-strength and nuisance-dimension sweeps use $n = 100$ labeled samples, $m = 20,000$ unlabeled samples, and $n_{\text{test}} = 10,000$ test samples. The unlabeled-sample-size sweep uses a smaller labeled budget, $n = 40$, while varying m .

Training procedure. All methods use linear predictors with an intercept. Baseline fits the deployment model f using only labeled pairs (X, Y) . Two-Stage first fits a rich-view teacher g using labeled pairs $((X, W), Y)$, pseudo-labels the unlabeled examples, and then fits an X -only student on the union of labeled and pseudo-labeled examples. Coupled Training solves the linear penalized objective for each value of λ in a logarithmic grid. For numerical stability, we use a small ridge penalty in all linear solves.

Hyperparameter selection. For the summary sweeps in the main text, λ is selected by cross-validation on the labeled set only. The unlabeled samples are used during training but their labels are never used for selecting λ . We sweep λ over a log-spaced grid from 10^{-4} to 10^4 .

Evaluation. For the controlled sweeps, we report the theory-aligned test estimation error

$$\mathbb{E}_{\text{test}}[(\hat{f}(X) - \mu(X))^2],$$

where $\mu(X) = X^\top \beta$ is known from the data-generating process. This removes irreducible label noise from the evaluation and directly measures how well the deployment predictor estimates the target regression function. We average results over multiple random seeds and report mean curves.

Controlled sweeps. The signal-strength sweep varies α while holding the nuisance dimension fixed. The nuisance-dimension sweep varies d_{noise} while holding the privileged signal strength fixed. The unlabeled-sample-size sweep varies m while keeping the labeled budget fixed. These three sweeps correspond respectively to the three panels in Figure 2, showing that Coupled Training improves as the privileged signal becomes useful, remains more stable than Two-Stage when nuisance privileged dimensions grow, and benefits from additional unlabeled paired (X, W) samples.

B.5. Synthetic Binary Classification Diagnostic

Dataset. We also run a controlled binary classification diagnostic to test the cross-entropy analogue of Coupled Training. The synthetic generator produces deployment features $X \in \mathbb{R}^5$ and privileged features $W \in \mathbb{R}^{40}$, with correlated deployment and privileged views. Labels are sampled from a Bernoulli model whose logit depends on both the deployment signal and the privileged signal, with additive logit noise. The default setting uses correlation strength 0.95, X scale 1.0, W scale 1.05, logit-noise standard deviation 0.70, and unlabeled mean parameter 1.0.

Feature split. The deployment view is $X \in \mathbb{R}^5$ and the privileged view is $W \in \mathbb{R}^{40}$. The final predictor must use only X at test time. The privileged block W is available only during training and is intentionally higher-dimensional than X , so the experiment tests whether Coupled Training can benefit from privileged information without fully inheriting rich-view noise.

Data split. For each random seed, we generate independent labeled, unlabeled, and test samples. We use

$$n = 50, \quad m = 3000, \quad n_{\text{test}} = 6000,$$

and average results over seeds $\{0, 1, 2, 3, 4\}$. The labels of the unlabeled samples are discarded during training and are used only by the data generator.

Models and hyperparameters. All methods use linear logistic predictors with intercepts. Let

$$p_f(x) = \sigma(\bar{x}^\top \beta), \quad p_g(x, w) = \sigma(\bar{z}^\top \gamma),$$

where $\bar{x}$ and $\bar{z}$ include intercepts and $\sigma(t) = (1 + e^{-t})^{-1}$. The deployment model uses ridge parameter $\alpha_f = 10^{-4}$, and the rich-view model uses ridge parameter $\alpha_g = 10^{-1}$. The X -only baseline is trained for 500 gradient steps with learning rate 0.05. The rich-view teacher and the Two-Stage student are each trained for 700 gradient steps with learning rate 0.03.

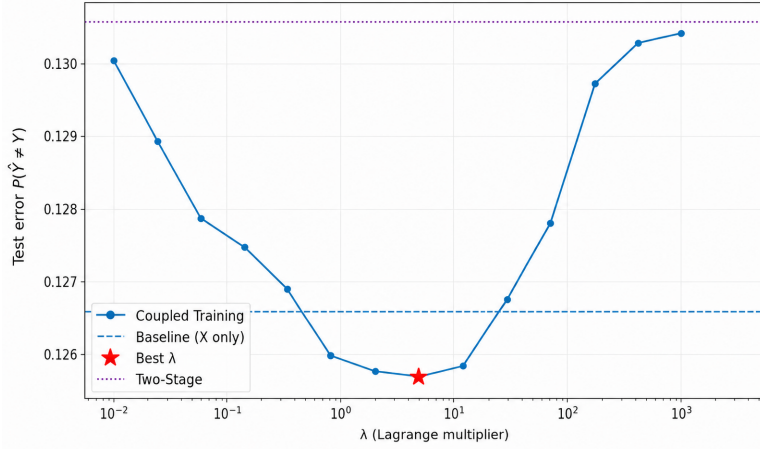


Figure 6. **Synthetic binary classification diagnostic.** Test 0–1 error for the cross-entropy analogue of Coupled Training as a function of λ , averaged over seeds $\{0, 1, 2, 3, 4\}$.

Training procedure. Baseline trains the deployment model using only labeled pairs (X_L, Y_L) . Two-Stage first trains a rich-view teacher using labeled pairs $((X_L, W_L), Y_L)$, pseudo-labels the unlabeled examples with teacher probabilities $p_g(X_U, W_U)$, and then trains an X -only student on the union of labeled and pseudo-labeled examples.

For Coupled Training, we use the cross-entropy analogue of the alternating updates. For $a \in [0, 1]$ and $p \in (0, 1)$, define

$$\text{CE}(a, p) = -a \log p - (1 - a) \log(1 - p).$$

Given g , the deployment update minimizes

$$\sum_{i \in L} \text{CE}(Y_i, p_f(X_i)) + \sum_{j \in U} \text{CE}(p_g(X_j, W_j), p_f(X_j)) + \frac{\alpha_f}{2} \|\beta_{-0}\|_2^2,$$

where the subscript -0 excludes the intercept. Given f , the rich-view update minimizes

$$\sum_{j \in U} \text{CE}(p_f(X_j), p_g(X_j, W_j)) + \lambda \sum_{i \in L} \text{CE}(Y_i, p_g(X_i, W_i)) + \frac{\alpha_g}{2} \|\gamma_{-0}\|_2^2.$$

We run 5 outer coupled iterations. Each f -update and each g -update uses 150 gradient steps with learning rate 0.02.

Hyperparameter sweep. For this diagnostic, λ is swept rather than selected by cross-validation. We use 14 log-spaced values from 10^{-2} to 10^3 . Thus, the best point in the figure should be interpreted as the best point in the displayed diagnostic sweep, not as a validation-selected tuning parameter.

Evaluation. The final test-time predictor is the deployment model $p_f(X)$, which uses only X . We classify by thresholding at $1/2$ and report test error

$$\mathbb{P}_{\text{test}}(\mathbf{1}\{p_f(X) \geq 1/2\} \neq Y).$$

All reported curves are averaged over the five random seeds.

Results. Figure 6 shows that the cross-entropy coupled model has an interior optimum in λ . The best coupled model occurs around $\lambda \approx 5$ and improves slightly over the labeled X -only baseline, while the Two-Stage limit performs worse. This mirrors the regression experiments, where moderate coupling can use privileged information, whereas overly strong transfer can propagate rich-view noise into the deployment model.

C. Proofs

In what follows, let $a = \frac{n}{N}$, $b = \frac{m}{N}$, and $c = b + a\lambda = \frac{m+n\lambda}{N}$, and define

$$s_f(z) := \frac{b f(x) + a\lambda \eta(z)}{c}, \quad t_g(x) := a\mu(x) + b\mathbb{E}[g(Z) \mid X = x].$$

For functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$, we define the (squared) L^2 norms $\|f\|_X^2 = \mathbb{E}[f^2(X)]$ and $\|g\|_Z^2 = \mathbb{E}[g^2(Z)]$. Define the inner products $\langle \cdot, \cdot \rangle_X$ and $\langle \cdot, \cdot \rangle_Z$ analogously.

For fixed points $s^n = (s_1, \dots, s_n)$ in a measurable space $\mathcal{S}$, let $\hat{P}_n$ denote the empirical measure with respect to these points. For a function $h : \mathcal{S} \rightarrow \mathbb{R}$, we write $\|h\|_{L^p(\hat{P}_n)} = (\frac{1}{n} \sum_{i=1}^n |h(s_i)|^p)^{1/p}$. An L^p - ϵ -cover of a class $\mathcal{C}$ with respect to $\|\cdot\|_{L^p(\hat{P}_n)}$ is a finite collection $\{h_1, \dots, h_M\} \subset \mathcal{C}$ such that every $h \in \mathcal{C}$ is within ϵ of some h_j under this norm, and the minimal such M is the empirical covering number $\mathcal{N}_p(\epsilon, \mathcal{C}, s^n)$.

C.1. Proof of Theorem 2.1

For a fixed $f \in \mathcal{F}$, let d denote constants that do not depend on g (and may change from line to line). Then, we can write the population loss as follows

$$\begin{aligned} \mathcal{L}(f, g; \lambda) &= b\mathbb{E}[(g(Z) - f(X))^2] + a\lambda\mathbb{E}[(Y - g(Z))^2] + d \\ &= b(\mathbb{E}g^2(Z) - 2\mathbb{E}g(Z)f(X) + \mathbb{E}f^2(X)) + a\lambda(\mathbb{E}Y^2 - 2\mathbb{E}[Yg(Z)] + \mathbb{E}g^2(Z)) + d \\ &= c\mathbb{E}g^2(Z) - 2\mathbb{E}[g(Z)(bf(X) + a\lambda Y)] + d \\ &= c\mathbb{E}g^2(Z) - 2\mathbb{E}[g(Z)\mathbb{E}[(bf(X) + a\lambda Y) \mid Z]] + d \\ &= c(\mathbb{E}g^2(Z) - 2\mathbb{E}[g(Z)s_f(Z)]) + d \\ &= c(\mathbb{E}[g^2(Z)] - 2\mathbb{E}[g(Z)s_f(Z)] + \mathbb{E}[s_f^2(Z)]) + d \end{aligned} \tag{5}$$

$$= c\|g - s_f\|_Z^2 + d. \tag{6}$$

Thus, for fixed $f \in \mathcal{F}$, the best response over $\mathcal{G}$ is the metric projection

$$g_{f,\lambda}^{\mathcal{G}} = \operatorname{argmin}_{g \in \mathcal{G}} \mathcal{L}(f, g; \lambda) = \operatorname{argmin}_{g \in \mathcal{G}} \|g - s_f\|_Z^2 = \Pi_{\mathcal{G}} s_f. \tag{7}$$

For a fixed $g \in \mathcal{G}$, let d denote constants that do not depend on f (and may change from line to line). Then using same approach to derive (6), we get

$$\begin{aligned} \mathcal{L}(f, g; \lambda) &= a\mathbb{E}[(Y - f(X))^2] + b\mathbb{E}[(g(Z) - f(X))^2] + d \\ &= (a + b)\mathbb{E}f^2(X) - 2\mathbb{E}[f(X)\mathbb{E}[(bg(Z) + aY) \mid X]] + d \\ &= \|f - t_g\|_X^2 + d. \end{aligned} \tag{8}$$

Thus, for fixed $g \in \mathcal{G}$, the best response over $\mathcal{F}$ is

$$f_g^{\mathcal{F}} = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f, g; \lambda) = \operatorname{argmin}_{f \in \mathcal{F}} \|f - t_g\|_X^2 = \Pi_{\mathcal{F}} t_g. \tag{9}$$

Hence, the desired result follows from (7) and (9).

Furthermore, if $\mu \in \mathcal{F}$ and $\eta \in \mathcal{G}$ (identifying μ with its lift $\tilde{\mu}(x, w) = \mu(x)$ on $\mathcal{Z}$), and $\eta \in \mathcal{G}$, set

$$f_0 := \mu, \quad g_0 := \frac{b}{c}\mu + \frac{a\lambda}{c}\eta.$$

The convexity of $\mathcal{G}$ gives $g_0 \in \mathcal{G}$, and $\mathbb{E}[g_0(Z) \mid X] = \mu(X)$. Consequently, $s_{f_0} = g_0$ and $t_{g_0} = f_0$. For any $(f, g) \in \mathcal{F} \times \mathcal{G}$, write $u = f - f_0$ and $v = g - g_0$. Using the conditional-mean identities for μ and η , together with $b(g_0 - f_0) = a\lambda(\eta - g_0)$, a direct expansion gives

$$\mathcal{L}(f, g; \lambda) - \mathcal{L}(f_0, g_0; \lambda) = a\|u\|_X^2 + b\|v - u\|_Z^2 + a\lambda\|v\|_Z^2 \geq 0.$$

Thus, (f_0, g_0) is a global minimizer. Since $a, b > 0$, equality forces $u = v = 0$, so the minimizer is unique and

$$f^* = \mu, \quad g^* = \frac{m}{m + n\lambda}\mu + \frac{n\lambda}{m + n\lambda}\eta.$$

Next, before proving Theorem 2.7, we present a sequence of intermediate results that will be useful for its proof.

Theorem C.1. *The following identity decomposes the excess population loss around (f^*, g^*) .*

$$\begin{aligned} \mathcal{L}(\hat{f}, \hat{g}; \lambda) - \mathcal{L}(f^*, g^*; \lambda) &= \|u\|_X^2 + c\|v\|_Z^2 - 2b\langle u, v \rangle_Z \\ &\quad + \langle \nabla_f \mathcal{L}(f^*, g^*; \lambda), u \rangle_X + \langle \nabla_g \mathcal{L}(f^*, g^*; \lambda), v \rangle_Z, \end{aligned} \quad (10)$$

where

$$u := \hat{f} - f^* \in \mathcal{F} - f^*, \quad v := \hat{g} - g^* \in \mathcal{G} - g^*.$$

Proof. We start by decomposing the loss $\mathcal{L}(\hat{f}, \hat{g}; \lambda)$ into three terms.

$$\mathcal{L}(\hat{f}, \hat{g}; \lambda) = \textcircled{1} + \textcircled{2} + \textcircled{3},$$

where

$$\textcircled{1} = a\mathbb{E}[(Y - \hat{f}(X))^2], \quad \textcircled{2} = b\mathbb{E}[(\hat{g}(Z) - \hat{f}(X))^2], \quad \text{and} \quad \textcircled{3} = a\lambda\mathbb{E}[(Y - \hat{g}(Z))^2].$$

Writing $\hat{f} = f^* + u$ and $\hat{g} = g^* + v$, we can expand each term as follows.

$$\begin{aligned} \textcircled{1} &= a(\mathbb{E}[(Y - f^*(X))^2] + \mathbb{E}u^2(X) - 2\mathbb{E}[(Y - f^*(X))u(X)]) \\ &= a\mathbb{E}[(Y - f^*(X))^2] + a\|u\|_X^2 - 2\mathbb{E}[u(X)(\mathbb{E}(Y | X) - f^*(X))] \\ &= a\mathbb{E}[(Y - f^*(X))^2] + a\|u\|_X^2 - 2a\langle \mu - f^*, u \rangle_X. \\ \textcircled{2} &= b(\mathbb{E}[(g^*(Z) - f^*(X))^2] + \mathbb{E}[(v(Z) - u(X))^2] + 2\mathbb{E}[(g^*(Z) - f^*(X))(v(Z) - u(X))]) \\ &= b\mathbb{E}[(g^*(Z) - f^*(X))^2] + b\|v\|_Z^2 + b\|u\|_X^2 - 2b\langle u, v \rangle_Z \\ &\quad + 2b\langle g^* - f^*, v \rangle_Z - 2b\langle \tilde{g}^* - f^*, u \rangle_X, \end{aligned}$$

where $\tilde{g}^*(x) = \mathbb{E}[g^*(Z) | X = x]$.

$$\begin{aligned} \textcircled{3} &= a\lambda\mathbb{E}[(Y - g^*(Z))^2] + a\lambda\|v\|_Z^2 - 2a\lambda\mathbb{E}[(Y - g^*(Z))v(Z)] \\ &= a\lambda\mathbb{E}[(Y - g^*(Z))^2] + a\lambda\|v\|_Z^2 - 2a\lambda\langle \eta - g^*, v \rangle_Z. \end{aligned}$$

Using these expansions, observe that

$$\begin{aligned} \mathcal{L}(\hat{f}, \hat{g}; \lambda) - \mathcal{L}(f^*, g^*; \lambda) &= (a + b)\|u\|_X^2 + c\|v\|_Z^2 - 2b\langle u, v \rangle_Z \\ &\quad + \langle 2(f^* - b\tilde{g}^* - a\mu), u \rangle_X + \langle 2(cg^* - bf^* - a\lambda\eta), v \rangle_Z \\ &= \|u\|_X^2 + c\|v\|_Z^2 - 2b\langle u, v \rangle_Z \\ &\quad + \langle 2(f^* - t_{g^*}), u \rangle_X + \langle 2c(g^* - s_{f^*}), v \rangle_Z \\ &= \|u\|_X^2 + c\|v\|_Z^2 - 2b\langle u, v \rangle_Z \\ &\quad + \langle \nabla_f \mathcal{L}(f^*, g^*; \lambda), u \rangle_X + \langle \nabla_g \mathcal{L}(f^*, g^*; \lambda), v \rangle_Z, \end{aligned}$$

where the last equality follows from (6) and (8), which give

$$\nabla_g \mathcal{L}(f, g; \lambda) = 2c(g - s_f), \quad \nabla_f \mathcal{L}(f, g; \lambda) = 2(f - t_g).$$

□

C.2. Proof of Corollary 2.3

Since (f^*, g^*) minimizes $\mathcal{L}$ over the closed convex set $\mathcal{F} \times \mathcal{G}$, the variational inequalities give

$$\langle \nabla_f \mathcal{L}(f^*, g^*; \lambda), u \rangle_X \geq 0, \quad \langle \nabla_g \mathcal{L}(f^*, g^*; \lambda), v \rangle_Z \geq 0$$

for all $u \in \mathcal{F} - f^*, v \in \mathcal{G} - g^*$.

Thus, dropping the nonnegative linear terms in (10) yields

$$\mathcal{L}(\hat{f}, \hat{g}; \lambda) - \mathcal{L}(f^*, g^*; \lambda) \geq \|u\|_X^2 + c\|v\|_Z^2 - 2b\langle u, v \rangle_Z. \quad (11)$$

The result follows by taking expectation with respect to $\mathcal{D}$ of both sides (11), then we can further lower bound the right hand by completing the square of the last two terms as done below.

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \|u\|_X^2 + c\mathbb{E}_{\mathcal{D}} \|v\|_Z^2 - 2b\mathbb{E}_{\mathcal{D}} \langle u, v \rangle_Z \\ & \geq \mathbb{E}_{\mathcal{D}} \|u\|_X^2 + c\mathbb{E}_{\mathcal{D}} \|v\|_Z^2 - 2b|\mathbb{E}_{\mathcal{D}} \langle u, v \rangle_Z| \\ & = \mathbb{E}_{\mathcal{D}} \|u\|_X^2 + c\mathbb{E}_{\mathcal{D}} \|v\|_Z^2 - 2b\rho_{\star} \sqrt{\mathbb{E}_{\mathcal{D}} \|u\|_X^2 \mathbb{E}_{\mathcal{D}} \|v\|_Z^2} \\ & = \mathbb{E}_{\mathcal{D}} \|u\|_X^2 + c \left(\mathbb{E}_{\mathcal{D}} \|v\|_Z^2 - \frac{2b\rho_{\star}}{c} \sqrt{\mathbb{E}_{\mathcal{D}} \|u\|_X^2 \mathbb{E}_{\mathcal{D}} \|v\|_Z^2} + \frac{b^2\rho_{\star}^2}{c^2} \mathbb{E}_{\mathcal{D}} \|u\|_X^2 \right) - \frac{b^2\rho_{\star}^2}{c} \mathbb{E}_{\mathcal{D}} \|u\|_X^2 \\ & = \gamma_{n,m,\lambda}(\rho_{\star}) \mathbb{E}_{\mathcal{D}} \|u\|_X^2 + c \left(\sqrt{\mathbb{E}_{\mathcal{D}} \|v\|_Z^2} - \frac{b\rho_{\star}}{c} \sqrt{\mathbb{E}_{\mathcal{D}} \|u\|_X^2} \right)^2 \\ & \geq \gamma_{n,m,\lambda}(\rho_{\star}) \mathbb{E}_{\mathcal{D}} \|u\|_X^2. \end{aligned}$$

Throughout, we use uppercase letters (e.g., X_i, Z_i, Y_i) for random variables and lowercase letters (e.g., x_i, z_i, y_i) for their realizations. We write $x^n = (x_1, \dots, x_n)$ and $z^n = (z_1, \dots, z_n)$ for the labeled sample locations, and $x^m = (x_{n+1}, \dots, x_N)$ and $z^m = (z_{n+1}, \dots, z_N)$ for the unlabeled ones (where $m = N - n$). We also use $x^N = (x_1, \dots, x_N)$ and $z^N = (z_1, \dots, z_N)$ to denote the full collection of sample points. Define

$$\begin{aligned} k_i(f, g) &:= (f(x_i) - y_i)^2 + \lambda(g(z_i) - y_i)^2 - [(f^*(x_i) - y_i)^2 + \lambda(g^*(z_i) - y_i)^2], \\ p_j(f, g) &:= (f(x_j) - g(z_j))^2 - (f^*(x_j) - g^*(z_j))^2. \end{aligned}$$

For notational convenience, throughout the remainder of the text we write $\sum_i$ for $\sum_{i=1}^n$ and $\sum_j$ for $\sum_{j=n+1}^N$. We now state a result adapted from Györfi et al. (2002, Theorem 11.4).

Theorem C.2. Assume that 2.5 holds.

Then

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \mathbb{E}[\hat{\mathcal{L}}(f, g; \lambda)] - \mathbb{E}[\hat{\mathcal{L}}(f^*, g^*; \lambda)] - \frac{1}{N} \left(\sum_i k_i + \sum_j p_j \right) \right. \\ & \quad \left. \geq \epsilon \left(\alpha + \beta + \mathbb{E}[\hat{\mathcal{L}}(f, g; \lambda)] - \mathbb{E}[\hat{\mathcal{L}}(f^*, g^*; \lambda)] \right) \right\} \\ & \leq 10\mathfrak{M}_{\mathcal{F}, \mathcal{G}}^{n,m,\lambda} \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha N}{512B^2(\lambda+1)} \right), \end{aligned}$$

where $\alpha, \beta > 0$, $0 < \epsilon < 1$, and

$$\begin{aligned} \mathfrak{M}_{\mathcal{F}, \mathcal{G}}^{n,m,\lambda} &:= \max_{z^n} \mathcal{N}_1 \left(\frac{\epsilon(\alpha + \beta)}{640Ba}, \mathcal{F}, x^n \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha + \beta)}{640B(\lambda+1)a}, \mathcal{G}, z^n \right) + \\ & \quad \max_{z^m} \mathcal{N}_1 \left(\frac{\epsilon(\alpha + \beta)}{640Bb}, \mathcal{F}, x^m \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha + \beta)}{640Bb}, \mathcal{G}, z^m \right) + \\ & \quad \max_{z^N} \mathcal{N}_1 \left(\frac{\epsilon\beta}{40B}, \mathcal{F}, x^N \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon\beta}{40B(\lambda+1)}, \mathcal{G}, z^N \right). \end{aligned}$$

Proof. Throughout this proof, for $(f, g) \in \mathcal{F} \times \mathcal{G}$, we interpret f and g pointwise, writing $f \equiv f(X)$, $g \equiv g(Z)$, so that expressions such as $(f - Y)^2$, $(g - Y)^2$, and $(f - g)^2$ are understood as $(f(X) - Y)^2$, $(g(Z) - Y)^2$, and $(f(X) - g(Z))^2$, respectively.

For a sample (X, Z, Y) , define the $*$ -norm

$$\|(f, g) - (Y, Y)\|_*^2 := a(f - Y)^2 + a\lambda(g - Y)^2 + b(f - g)^2.$$

Denote

$$A_{f,g} := \|(f, g) - (Y, Y)\|_*^2 - \|(f^*, g^*) - (Y, Y)\|_*^2 = a k(f, g) + b p(f, g),$$

where $k(f, g) = (f - Y)^2 + \lambda(g - Y)^2 - [(f^* - Y)^2 + \lambda(g^* - Y)^2]$, and $p(f, g) = (f - g)^2 - (f^* - g^*)^2$.
By the definition above, our goal is to bound

$$\mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \mathbb{E}[A_{f,g}] - \frac{1}{N} \sum_i k_i - \frac{1}{N} \sum_j p_j \geq \epsilon(\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \quad (12)$$

Observe that when $N \leq \frac{128(\lambda + 1)B^2}{\epsilon^2(\alpha + \beta)}$, we have

$$10\mathfrak{M}_{\mathcal{F}, \mathcal{G}}^{n,m,\lambda} \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha N}{512B^2(\lambda+1)} \right) \geq 15 \exp \left(-\frac{\epsilon^2(1-\epsilon)(\alpha + \beta)N}{512B^2(\lambda+1)} \right) \geq 15 \exp \left(-\frac{128}{512} \right) > 1,$$

thus it is sufficient to only consider the case when $N > \frac{128(\lambda + 1)B^2}{\epsilon^2(\alpha + \beta)}$.

To bound (12), we divide the proof into several steps.

Step 1. Symmetrization

Let $\mathcal{D}'_L = \{(X'_i, W'_i, Y'_i)\}_{i=1}^n \stackrel{\text{i.i.d.}}{\sim} P_{X,W,Y}$, and $\mathcal{D}'_U = \{Z'_j = (X'_j, W'_j)\}_{j=n+1}^N \stackrel{\text{i.i.d.}}{\sim} P_{X,W}$, be a *ghost* sample independent of $\mathcal{D}_L$ and $\mathcal{D}_U$, respectively. Denote

$$\begin{aligned} k'_i &= (f(X'_i) - Y'_i)^2 + \lambda(g(Z'_i) - Y'_i)^2 - (f^*(X'_i) - Y'_i)^2 - \lambda(g^*(Z'_i) - Y'_i)^2, \\ p'_j &= (f(X'_j) - g(Z'_j))^2 - (f^*(X'_j) - g^*(Z'_j))^2, \end{aligned}$$

and define the original and ghost collections of labeled and unlabeled summands

$$\mathcal{D} = \{k_i, p_j : 1 \leq i \leq n, n+1 \leq j \leq N\}, \quad \mathcal{D}' = \{k'_i, p'_j : 1 \leq i \leq n, n+1 \leq j \leq N\}.$$

We will show that

$$\begin{aligned} &\mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \mathbb{E}[A_{f,g}] - \frac{1}{N} \sum_i k_i - \frac{1}{N} \sum_j p_j \geq \epsilon(\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \\ &\leq \frac{8}{7} \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2}(\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \end{aligned} \quad (13)$$

To prove this, consider

$$\begin{aligned}
 \text{Var} \left(\frac{1}{N} \left(\sum_i k'_i + \sum_j p'_j \right) \right) &= \frac{1}{N^2} \text{Var} \left(\sum_i k'_i + \sum_j p'_j \right) \leq \frac{1}{N^2} (n \text{Var}(k'_1) + m \text{Var}(p'_1)) \\
 &\leq \frac{a}{N} \mathbb{E} \left[\left((f-y)^2 + \lambda(g-y)^2 - (f^*-y)^2 - \lambda(g^*-y)^2 \right)^2 \right] + \frac{b}{N} \mathbb{E} \left[\left((f-g)^2 - (f^*-g^*)^2 \right)^2 \right] \\
 &\leq \frac{1}{N} \mathbb{E} \left[a((f+f^*-2y)(f-f^*) + \lambda(g+g^*-2y)(g-g^*))^2 \right. \\
 &\quad \left. + b((f-g+f^*-g^*)(f-g-f^*+g^*))^2 \right] \\
 &\stackrel{(*)}{\leq} \frac{1}{N} \mathbb{E} \left[a((f+f^*-2y)^2 + \lambda(g+g^*-2y)^2)((f-f^*)^2 + \lambda(g-g^*)^2) \right. \\
 &\quad \left. + b((f-g+f^*-g^*)(f-g-f^*+g^*))^2 \right] \\
 &\leq \frac{16(\lambda+1)B^2}{N} \mathbb{E} \left[(a((f-f^*)^2 + \lambda(g-g^*)^2) + b(f-f^*+g^*-g)^2) \right] \\
 &= \frac{16(\lambda+1)B^2}{N} \mathbb{E} \left[\|(f, g) - (f^*, g^*)\|_*^2 \right] \\
 &\stackrel{(\dagger)}{\leq} \frac{16(\lambda+1)B^2}{N} \mathbb{E} \left[\|(f, g) - (Y, Y)\|_*^2 - \|(f^*, g^*) - (Y, Y)\|_*^2 \right] \\
 &\leq \frac{16(\lambda+1)B^2}{N} \mathbb{E} [A_{f,g}], \tag{14}
 \end{aligned}$$

where $(*)$ follows by applying Cauchy–Schwarz, and $(\dagger)$ holds since (f^*, g^*) minimizes $\mathbb{E} \|(f, g) - (Y, Y)\|_*^2 = \mathcal{L}(f, g; \lambda)$. Thus,

$$\mathbb{E} [A_{f,g}] \geq \frac{1}{16(\lambda+1)B^2} (a\mathbb{E}[k^2(f, g)] + b\mathbb{E}[p^2(f, g)]) \tag{15}$$

Let $\mathcal{E}$ denote the event on the left-hand side of (13). For every fixed $(f, g) \in \mathcal{F} \times \mathcal{G}$, Chebyshev’s inequality and (14) give

$$\begin{aligned}
 &\mathbb{P}_{\mathcal{D}'} \left\{ \mathbb{E}[A_{f,g}] - \frac{1}{N} \left(\sum_i k'_i + \sum_j p'_j \right) > \frac{\epsilon}{2} (\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \\
 &\leq \frac{64(\lambda+1)B^2 \mathbb{E}[A_{f,g}]}{N\epsilon^2 (\alpha + \beta + \mathbb{E}[A_{f,g}])^2} \\
 &\leq \frac{16(\lambda+1)B^2}{N\epsilon^2 (\alpha + \beta)} < \frac{1}{8},
 \end{aligned}$$

where the last two inequalities use $\mathbb{E}[A_{f,g}] \geq 0$ and the lower bound on N above.

Let $\mathcal{H}$ be the event that there exists the same $(f, g) \in \mathcal{F} \times \mathcal{G}$ satisfying both

$$\mathbb{E}[A_{f,g}] - \frac{1}{N} \left(\sum_i k_i + \sum_j p_j \right) \geq \epsilon (\alpha + \beta + \mathbb{E}[A_{f,g}])$$

and

$$\mathbb{E}[A_{f,g}] - \frac{1}{N} \left(\sum_i k'_i + \sum_j p'_j \right) \leq \frac{\epsilon}{2} (\alpha + \beta + \mathbb{E}[A_{f,g}]).$$

For every realization $\mathcal{D} \in \mathcal{E}$, the conditional section $\mathcal{H}_{\mathcal{D}}$ contains the ghost-good event corresponding to any witness of $\mathcal{E}$. Hence

$$\mathbb{P}(\mathcal{H} \mid \mathcal{D}) \geq \frac{7}{8} \quad \text{on } \mathcal{E}, \quad \mathbb{P}(\mathcal{H}) \geq \frac{7}{8} \mathbb{P}(\mathcal{E}).$$

On $\mathcal{H}$, the same pair (f, g) satisfies

$$\frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta + \mathbb{E}[A_{f,g}]).$$

Therefore (13) follows.

Step 2. Population to empirical

Define

$$\widehat{Q}(f, g) := \frac{1}{N} \left(\sum_i k_i^2 + \sum_j p_j^2 \right), \quad \widehat{Q}'(f, g) := \frac{1}{N} \left(\sum_i (k'_i)^2 + \sum_j (p'_j)^2 \right),$$

and set

$$Q(f, g) := \mathbb{E}[ak^2 + bp^2], \quad \tau := B^2(\lambda + 1)(\alpha + \beta) = \tau_\alpha + \tau_\beta, \quad \tau_\alpha := B^2(\lambda + 1)\alpha, \quad \tau_\beta := B^2(\lambda + 1)\beta.$$

Next, we show that

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \\ & \leq \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta) \right. \\ & \quad \left. + \frac{\epsilon}{64(\lambda + 1)B^2} \frac{1 - \epsilon}{1 + \epsilon} \left(\widehat{Q}(f, g) + \widehat{Q}'(f, g) - \frac{2\epsilon}{1 - \epsilon} \tau \right) \right\} \\ & + 8\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\epsilon\tau}{80aB^2(\lambda + 1)}, \mathcal{G}_k, z^n \right) + \mathcal{N}_1 \left(\frac{\epsilon\tau}{80bB^2(\lambda + 1)}, \mathcal{G}_p, z^m \right) \right] \exp \left(-\frac{\epsilon^2\tau N}{480B^4(\lambda + 1)^2} \right). \end{aligned} \quad (16)$$

To prove (16), we first present the following lemma.

Lemma C.3. *Let $\mathcal{G}_k = \{k(f, g) : (f, g) \in \mathcal{F} \times \mathcal{G}\}$, $\mathcal{G}_p = \{p(f, g) : (f, g) \in \mathcal{F} \times \mathcal{G}\}$. $(z^n, y^n), z^m$ are the samples used to generate k_i, p_j respectively. Then*

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{\widehat{Q}(f, g) - Q(f, g)}{\tau + \widehat{Q}(f, g) + Q(f, g)} > \epsilon \right\} \\ & \leq 4\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\epsilon\tau}{80aB^2(\lambda + 1)}, \mathcal{G}_k, z^n \right) + \mathcal{N}_1 \left(\frac{\epsilon\tau}{80bB^2(\lambda + 1)}, \mathcal{G}_p, z^m \right) \right] \exp \left(-\frac{\epsilon^2\tau N}{480B^4(\lambda + 1)^2} \right). \end{aligned}$$

Proof. Write

$$P_n k^2 := \frac{1}{n} \sum_i k_i^2, \quad P_m p^2 := \frac{1}{m} \sum_j p_j^2.$$

Since $\widehat{Q} = aP_n k^2 + bP_m p^2$ and $Q = a\mathbb{E}[k^2] + b\mathbb{E}[p^2]$, the event in the lemma is contained in the union of

$$\begin{aligned} & \left\{ \exists (f, g) : \frac{P_n k^2 - \mathbb{E}[k^2]}{\tau/(2a) + P_n k^2 + \mathbb{E}[k^2]} > \epsilon \right\}, \\ & \left\{ \exists (f, g) : \frac{P_m p^2 - \mathbb{E}[p^2]}{\tau/(2b) + P_m p^2 + \mathbb{E}[p^2]} > \epsilon \right\}. \end{aligned}$$

Indeed, if neither block event occurs, multiplying the two inequalities by a and b and adding them gives $\widehat{Q} - Q \leq \epsilon(\tau + \widehat{Q} + Q)$. Introduce the nonnegative squared classes

$$\mathcal{H}_k := \{k^2 : k \in \mathcal{G}_k\}, \quad \mathcal{H}_p := \{p^2 : p \in \mathcal{G}_p\}.$$

Under Assumption 2.5,

$$|k| \leq 4B^2(\lambda + 1), \quad |p| \leq 4B^2,$$

and hence

$$|k^2 - \tilde{k}^2| \leq 8B^2(\lambda + 1)|k - \tilde{k}|, \quad |p^2 - \tilde{p}^2| \leq 8B^2|p - \tilde{p}|.$$

The two squared classes have envelopes $16B^4(\lambda + 1)^2$ and $16B^4$, respectively. For a nonnegative class $\mathcal{H}$ bounded by L , an i.i.d. sample of size r , and $\nu > 0$, Györfi et al. (2002, Theorem 11.6) gives

$$\begin{aligned} \mathbb{P} \left\{ \exists h \in \mathcal{H} : \frac{P_r h - \mathbb{E}h}{\nu + P_r h + \mathbb{E}h} > \epsilon \right\} \\ \leq 4\mathbb{E}\mathcal{N}_1 \left(\frac{\epsilon\nu}{5}, \mathcal{H}, S^r \right) \exp \left(-\frac{3\epsilon^2\nu r}{40L} \right). \end{aligned}$$

If necessary, this form follows by rescaling the class. Apply it to $\mathcal{H}_k$ with $\nu = \tau/(2a)$ and to $\mathcal{H}_p$ with $\nu = \tau/(2b)$. Since $n = aN$ and $m = bN$, the lemma probability is at most

$$\begin{aligned} 4\mathbb{E}\mathcal{N}_1 \left(\frac{\epsilon\tau}{10a}, \mathcal{H}_k, z^n \right) \exp \left(-\frac{3\epsilon^2\tau N}{1280B^4(\lambda + 1)^2} \right) \\ + 4\mathbb{E}\mathcal{N}_1 \left(\frac{\epsilon\tau}{10b}, \mathcal{H}_p, z^m \right) \exp \left(-\frac{3\epsilon^2\tau N}{1280B^4} \right). \end{aligned}$$

The preceding Lipschitz bounds pull these squared-class covers back to $\mathcal{G}_k$ and $\mathcal{G}_p$. Since $\lambda + 1 \geq 1$ and $3/1280 \geq 1/480$, weakening both terms to a common radius and exponent gives

$$\begin{aligned} \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{\widehat{Q}(f, g) - Q(f, g)}{\tau + \widehat{Q}(f, g) + Q(f, g)} > \epsilon \right\} \\ \leq 4\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\epsilon\tau}{80aB^2(\lambda + 1)}, \mathcal{G}_k, z^n \right) + \mathcal{N}_1 \left(\frac{\epsilon\tau}{80bB^2(\lambda + 1)}, \mathcal{G}_p, z^m \right) \right] \exp \left(-\frac{\epsilon^2\tau N}{480B^4(\lambda + 1)^2} \right). \end{aligned}$$

□

Next, define the events

$$\begin{aligned} \mathcal{S}_1 &:= \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\}, \\ \mathcal{S}_2 &:= \left\{ \forall (f, g) \in \mathcal{F} \times \mathcal{G} : \widehat{Q}(f, g) - Q(f, g) \leq \epsilon(\tau + \widehat{Q}(f, g) + Q(f, g)) \right\}, \\ \mathcal{S}_3 &:= \left\{ \forall (f, g) \in \mathcal{F} \times \mathcal{G} : \widehat{Q}'(f, g) - Q(f, g) \leq \epsilon(\tau + \widehat{Q}'(f, g) + Q(f, g)) \right\}. \end{aligned}$$

We have the set inclusion $\mathcal{S}_1 \subset (\mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_3) \cup \mathcal{S}_2^c \cup \mathcal{S}_3^c$, hence

$$\begin{aligned} \mathbb{P}(\mathcal{S}_1) &\leq \mathbb{P}(\mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_3) + \mathbb{P}(\mathcal{S}_2^c) + \mathbb{P}(\mathcal{S}_3^c) \\ &\leq \mathbb{P}(\mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_3) + 2\mathbb{P}(\mathcal{S}_2^c) \\ &\leq \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta) \right. \\ &\quad \left. + \frac{\epsilon}{64(\lambda + 1)B^2} \frac{1 - \epsilon}{1 + \epsilon} \left(\widehat{Q}(f, g) + \widehat{Q}'(f, g) - \frac{2\epsilon}{1 - \epsilon} \tau \right) \right\} \\ &\quad + 8\mathbb{E} \left[\mathcal{N}_1 \left(\frac{\epsilon\tau}{80aB^2(\lambda + 1)}, \mathcal{G}_k, z^n \right) + \mathcal{N}_1 \left(\frac{\epsilon\tau}{80bB^2(\lambda + 1)}, \mathcal{G}_p, z^m \right) \right] \exp \left(-\frac{\epsilon^2\tau N}{480B^4(\lambda + 1)^2} \right), \end{aligned}$$

where the last inequality uses Lemma C.3 and the fact that $\mathcal{S}_2$ implies

$$\mathbb{E}[A_{f,g}] \stackrel{(*)}{\geq} \frac{1}{16(\lambda+1)B^2} Q(f,g) \geq \frac{1}{16(\lambda+1)B^2} \frac{1-\epsilon}{1+\epsilon} \left(\widehat{Q}(f,g) - \frac{\epsilon}{1-\epsilon} \tau \right)$$

Similarly for $\mathcal{D}'$, $\mathcal{S}_3$ implies

$$\mathbb{E}[A_{f,g}] \stackrel{(*)}{\geq} \frac{1}{16(\lambda+1)B^2} Q(f,g) \geq \frac{1}{16(\lambda+1)B^2} \frac{1-\epsilon}{1+\epsilon} \left(\widehat{Q}'(f,g) - \frac{\epsilon}{1-\epsilon} \tau \right),$$

where $(*)$ follows by (15). Therefore (16) holds.

Step 3. Rademacher variables. Let $\{U_\ell\}_{\ell=1}^N$ be independent Rademacher variables, independent of both the original and ghost samples. Set $c_\epsilon := \frac{\epsilon}{64B^2(\lambda+1)} \frac{1-\epsilon}{1+\epsilon}$. Then

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f,g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left(\sum_i (k'_i - k_i) + \sum_j (p'_j - p_j) \right) \geq \frac{\epsilon}{2} (\alpha + \beta) \right. \\ & \quad \left. + c_\epsilon \left(\widehat{Q}(f,g) + \widehat{Q}'(f,g) - \frac{2\epsilon}{1-\epsilon} \tau \right) \right\} \\ & \leq 2\mathbb{P} \left\{ \exists (f,g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left| \sum_i U_i k_i + \sum_j U_j p_j \right| \geq \frac{\epsilon}{4} (\alpha + \beta) \right. \\ & \quad \left. + c_\epsilon \left(\widehat{Q}(f,g) - \frac{\epsilon}{1-\epsilon} \tau \right) \right\}, \end{aligned} \tag{17}$$

For brevity, write

$$\begin{aligned} R_{\mathcal{D}}(f,g) &:= \frac{\epsilon}{4} (\alpha + \beta) + c_\epsilon \left(\widehat{Q}(f,g) - \frac{\epsilon}{1-\epsilon} \tau \right), \\ T_{\mathcal{D}}(f,g;U) &:= \frac{1}{N} \left(\sum_i U_i k_i + \sum_j U_j p_j \right), \end{aligned}$$

and define $R_{\mathcal{D}'}$ and $T_{\mathcal{D}'}$ analogously. Independently swapping the two observations within each original–ghost pair shows that the left-hand probability in (17) equals

$$\mathbb{P} \{ \exists (f,g) : T_{\mathcal{D}'}(f,g;U) - T_{\mathcal{D}}(f,g;U) \geq R_{\mathcal{D}}(f,g) + R_{\mathcal{D}'}(f,g) \},$$

because $\widehat{Q} + \widehat{Q}'$ is invariant under these swaps. By the triangle inequality and a union bound, this is at most

$$\begin{aligned} & \mathbb{P} \{ \exists (f,g) : |T_{\mathcal{D}}(f,g;U)| \geq R_{\mathcal{D}}(f,g) \} \\ & + \mathbb{P} \{ \exists (f,g) : |T_{\mathcal{D}'}(f,g;U)| \geq R_{\mathcal{D}'}(f,g) \}. \end{aligned}$$

The two probabilities are equal by identical distribution, which proves (17).

Step 4. Covering numbers. For each realization $\mathcal{D}$, let $\Gamma_\delta = \Gamma_\delta(\mathcal{D})$ be a smallest subset of $\mathcal{F} \times \mathcal{G}$ such that, for every $(f,g) \in \mathcal{F} \times \mathcal{G}$, there exists $(\tilde{f}, \tilde{g}) \in \Gamma_\delta$ satisfying

$$\frac{1}{N} \left(\sum_i |k_i(f,g) - k_i(\tilde{f}, \tilde{g})| + \sum_j |p_j(f,g) - p_j(\tilde{f}, \tilde{g})| \right) \leq \delta.$$

For notational simplicity, we write $\tilde{k}_i$ and $\tilde{p}_j$ in place of $k_i(\tilde{f}, \tilde{g})$ and $p_j(\tilde{f}, \tilde{g})$, respectively. Define

$$\tilde{Q} := \frac{1}{N} \left(\sum_i \tilde{k}_i^2 + \sum_j \tilde{p}_j^2 \right).$$

Conditionally on $\mathcal{D}$, we show that

$$\begin{aligned} & \mathbb{P}_U \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left| \sum_i U_i k_i + \sum_j U_j p_j \right| \geq \frac{\epsilon}{4} (\alpha + \beta) + c_\epsilon \left(\hat{Q}(f, g) - \frac{\epsilon}{1-\epsilon} \tau \right) \mid \mathcal{D} \right\} \\ & \leq \left| \Gamma_{\frac{\epsilon\beta}{5}} \right| \max_{(\tilde{f}, \tilde{g}) \in \Gamma_{\frac{\epsilon\beta}{5}}} \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \geq \frac{\epsilon}{4} \alpha + c_\epsilon \left(\tilde{Q} - \frac{\epsilon}{1-\epsilon} \tau_\alpha \right) \mid \mathcal{D} \right\}. \end{aligned} \quad (18)$$

Indeed, conditionally on $\mathcal{D}$, we have

$$\begin{aligned} & \mathbb{P}_U \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left| \sum_i U_i k_i + \sum_j U_j p_j \right| \geq \frac{\epsilon}{4} (\alpha + \beta) + c_\epsilon \left(\hat{Q}(f, g) - \frac{\epsilon}{1-\epsilon} \tau \right) \mid \mathcal{D} \right\} \\ & \stackrel{(*)}{\leq} |\Gamma_\delta| \max_{(\tilde{f}, \tilde{g}) \in \Gamma_\delta} \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| + \delta \geq \frac{\epsilon}{4} (\alpha + \beta) + c_\epsilon \left(\tilde{Q} - 8B^2(\lambda + 1)\delta - \frac{\epsilon}{1-\epsilon} \tau \right) \mid \mathcal{D} \right\} \\ & \stackrel{(\dagger)}{\leq} \left| \Gamma_{\frac{\epsilon\beta}{5}} \right| \max_{(\tilde{f}, \tilde{g}) \in \Gamma_{\frac{\epsilon\beta}{5}}} \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \geq \frac{\epsilon}{4} \alpha + c_\epsilon \left(\tilde{Q} - \frac{\epsilon}{1-\epsilon} \tau_\alpha \right) \mid \mathcal{D} \right\}, \end{aligned}$$

where $(*)$ follows by observing that

$$\begin{aligned} & \frac{1}{N} \left| \sum_i U_i k_i + \sum_j U_j p_j \right| - \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \leq \frac{1}{N} \left| \sum_i U_i (k_i - \tilde{k}_i) + \sum_j U_j (p_j - \tilde{p}_j) \right| \\ & \leq \frac{1}{N} \left(\sum_i |k_i(f, g) - k_i(\tilde{f}, \tilde{g})| + \sum_j |p_j(f, g) - p_j(\tilde{f}, \tilde{g})| \right) \leq \delta, \end{aligned}$$

and

$$\begin{aligned} & \frac{1}{N} \left(\sum_i \tilde{k}_i^2 + \sum_j \tilde{p}_j^2 \right) - \frac{1}{N} \left(\sum_i k_i^2 + \sum_j p_j^2 \right) \\ & \leq \frac{1}{N} \left(\sum_i (\tilde{k}_i - k_i)(\tilde{k}_i + k_i) + \sum_j (\tilde{p}_j - p_j)(\tilde{p}_j + p_j) \right) \\ & \leq \frac{8B^2(\lambda + 1)}{N} \left(\sum_i |k_i(f, g) - k_i(\tilde{f}, \tilde{g})| + \sum_j |p_j(f, g) - p_j(\tilde{f}, \tilde{g})| \right) \leq 8B^2(\lambda + 1)\delta. \end{aligned}$$

$(\dagger)$ follows by setting $\delta = \frac{\epsilon\beta}{5}$, using $\tau = \tau_\alpha + \tau_\beta$, and observing that

$$\begin{aligned} & \frac{\epsilon\beta}{4} - \frac{\epsilon\beta}{5} - c_\epsilon \left(8B^2(\lambda + 1) \frac{\epsilon\beta}{5} + \frac{\epsilon}{1-\epsilon} \tau_\beta \right) \\ & = \frac{\epsilon\beta}{20} - \frac{\epsilon^2(1-\epsilon)}{40(1+\epsilon)}\beta - \frac{\epsilon^2}{64(1+\epsilon)}\beta \geq 0. \end{aligned}$$

Step 5. Bernstein's inequality. Conditionally on $\mathcal{D}$, for every fixed $(\tilde{f}, \tilde{g}) \in \Gamma_{\epsilon\beta/5}(\mathcal{D})$, Bernstein's inequality gives

$$\begin{aligned} & \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \geq \frac{\epsilon}{4} \alpha + c_\epsilon \left(\tilde{Q} - \frac{\epsilon}{1-\epsilon} \tau_\alpha \right) \mid \mathcal{D} \right\} \\ & \leq 2 \exp \left(- \frac{N\epsilon^2(1-\epsilon)\alpha}{512(\lambda + 1)B^2} \right). \end{aligned} \quad (19)$$

To verify this bound, define

$$\sigma^2 := \frac{1}{N} \left(\sum_i \tilde{k}_i^2 + \sum_j \tilde{p}_j^2 \right) = \frac{1}{N} \text{Var} \left(\sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right),$$

$$A_1 := \frac{\epsilon}{4} \alpha - \frac{\epsilon^2}{64(1+\epsilon)} \alpha, \text{ and } A_2 := \frac{\epsilon}{64(\lambda+1)B^2} \frac{1-\epsilon}{1+\epsilon}.$$

By Bernstein's inequality and the independence of the Rademacher variables, we have

$$\begin{aligned} & \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \geq \frac{\epsilon}{4} \alpha + c_\epsilon \left(\tilde{Q} - \frac{\epsilon}{1-\epsilon} \tau_\alpha \right) \mid \mathcal{D} \right\} \\ &= \mathbb{P}_U \left\{ \frac{1}{N} \left| \sum_i U_i \tilde{k}_i + \sum_j U_j \tilde{p}_j \right| \geq A_1 + A_2 \sigma^2 \mid \mathcal{D} \right\} \\ &\leq 2 \exp \left(- \frac{N(A_1 + A_2 \sigma^2)^2}{2\sigma^2 + \frac{8}{3} B^2 (\lambda+1) (A_1 + A_2 \sigma^2)} \right) = 2 \exp \left(- \frac{3NA_2}{8B^2(\lambda+1)} \cdot \frac{(A_1/A_2 + \sigma^2)^2}{\left(\frac{3}{4B^2 A_2 (\lambda+1)} + 1 \right) \sigma^2 + \frac{A_1}{A_2}} \right) \\ &\stackrel{(\dagger)}{\leq} 2 \exp \left(- \frac{3NA_2}{8B^2(\lambda+1)} \cdot \frac{192 \epsilon (1+\epsilon) (\lambda+1) B^2 (16+15\epsilon) \alpha}{(\epsilon(1-\epsilon) + 48(1+\epsilon))^2} \right) \\ &\leq 2 \exp \left(- \frac{3NA_2}{8B^2(\lambda+1)} \cdot \frac{192 \epsilon (1+\epsilon) 16(\lambda+1) B^2 \alpha}{(\epsilon(1-\epsilon) + 48(1+\epsilon))^2} \right) \\ &= 2 \exp \left(- \frac{3N}{8B^2(\lambda+1)} \cdot \frac{\epsilon(1-\epsilon)}{64(\lambda+1)B^2(1+\epsilon)} \cdot \frac{192 \epsilon (1+\epsilon) 16(\lambda+1) B^2 \alpha}{(\epsilon(1-\epsilon) + 48(1+\epsilon))^2} \right) \\ &= 2 \exp \left(- \frac{18N \epsilon^2 (1-\epsilon) \alpha}{(\lambda+1) B^2 (\epsilon(1-\epsilon) + 48(1+\epsilon))^2} \right) \leq 2 \exp \left(- \frac{N \epsilon^2 (1-\epsilon) \alpha}{512(\lambda+1) B^2} \right), \end{aligned}$$

where the last inequality follows using $\max_{0 \leq \epsilon \leq 1} (\epsilon(1-\epsilon) + 48(1+\epsilon))^2 = 9216$. $(\dagger)$ follows by noting that for any $u > 0$, $b > 2$, we have

$$\frac{(a+u)^2}{a+bu} \geq \frac{(a + \frac{b-2}{b}a)^2}{a + b \frac{b-2}{b}a} = 4a \frac{b-1}{b^2},$$

which yields that,

$$\begin{aligned} \frac{(A_1/A_2 + \sigma^2)^2}{\left(\frac{3}{4B^2 A_2 (\lambda+1)} + 1 \right) \sigma^2 + \frac{A_1}{A_2}} &\geq \frac{\frac{192A_1(1+\epsilon)}{\epsilon(1-\epsilon)}}{A_2 \left(1 + \frac{48(1+\epsilon)}{\epsilon(1-\epsilon)} \right)^2} = \frac{192 \alpha \epsilon (1+\epsilon) (\lambda+1) B^2 (16+15\epsilon)}{(\epsilon(1-\epsilon) + 48(1+\epsilon))^2} \\ &\geq \frac{192 \epsilon (1+\epsilon) 16(\lambda+1) B^2 \alpha}{(\epsilon(1-\epsilon) + 48(1+\epsilon))^2}. \end{aligned}$$

Therefore (19) holds.

Averaging (18) over $\mathcal{D}$ and using (19), we obtain

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \frac{1}{N} \left| \sum_i U_i k_i + \sum_j U_j p_j \right| \geq \frac{\epsilon}{4} (\alpha + \beta) + c_\epsilon \left(\hat{Q}(f, g) - \frac{\epsilon}{1-\epsilon} \tau \right) \right\} \\ &\leq 2 \mathbb{E}_{\mathcal{D}} |\Gamma_{\epsilon\beta/5}(\mathcal{D})| \exp \left(- \frac{N \epsilon^2 (1-\epsilon) \alpha}{512(\lambda+1) B^2} \right). \end{aligned} \tag{20}$$

Step 6. Final bound. Observe that if we have

$$\begin{aligned} \frac{1}{N} \left(\sum_i |f(X_i) - \tilde{f}(X_i)| + \sum_j |f(X_j) - \tilde{f}(X_j)| \right) &\leq \frac{\epsilon\beta}{40B}, \\ \frac{1}{N} \left(\sum_i |g(Z_i) - \tilde{g}(Z_i)| + \sum_j |g(Z_j) - \tilde{g}(Z_j)| \right) &\leq \frac{\epsilon\beta}{40B(\lambda+1)}, \end{aligned}$$

then we have

$$\begin{aligned} &\frac{1}{N} \left(\sum_i |k_i(f, g) - k_i(\tilde{f}, \tilde{g})| + \sum_j |p_j(f, g) - p_j(\tilde{f}, \tilde{g})| \right) \\ &\leq \frac{1}{N} \left(\sum_i 4B |f(X_i) - \tilde{f}(X_i)| + 4\lambda B |g(Z_i) - \tilde{g}(Z_i)| + \sum_j 4B |f(X_j) - \tilde{f}(X_j)| + 4B |g(Z_j) - \tilde{g}(Z_j)| \right) \leq \frac{\epsilon\beta}{5}, \end{aligned}$$

which implies that $|\Gamma_{\frac{\epsilon\beta}{5}}| \leq \mathcal{N}_1 \left(\frac{\epsilon\beta}{40B}, \mathcal{F}, x^N \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon\beta}{40(\lambda+1)B}, \mathcal{G}, z^N \right)$. Thus $\mathbb{E}_{\mathcal{D}} |\Gamma_{\epsilon\beta/5}(\mathcal{D})| \leq \mathcal{M}_1$, with $\mathcal{M}_1$ defined below.

Combining the steps, we have the following.

$$\begin{aligned} &\mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \mathbb{E}[A_{f,g}] - \frac{1}{N} \sum_i k_i - \frac{1}{N} \sum_j p_j \geq \epsilon(\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \\ &\leq \frac{32}{7} \mathcal{M}_1 \exp \left(-\frac{N\epsilon^2(1-\epsilon)\alpha}{512(\lambda+1)B^2} \right) + \frac{64}{7} \mathcal{M}_2 \exp \left(-\frac{\epsilon^2\tau N}{480B^4(\lambda+1)^2} \right) \end{aligned}$$

where $\mathcal{M}_1 = \max_{z^N} \mathcal{N}_1 \left(\frac{\epsilon\beta}{40B}, \mathcal{F}, x^N \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon\beta}{40(\lambda+1)B}, \mathcal{G}, z^N \right)$, and

$$\begin{aligned} \mathcal{M}_2 &= \max_{z^n} \mathcal{N}_1 \left(\frac{\epsilon\tau}{80aB^2(\lambda+1)}, \mathcal{G}_k, z^n \right) + \max_{z^m} \mathcal{N}_1 \left(\frac{\epsilon\tau}{80bB^2(\lambda+1)}, \mathcal{G}_p, z^m \right) \\ &\leq \max_{z^n} \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Ba}, \mathcal{F}, x^n \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640B(\lambda+1)a}, \mathcal{G}, z^n \right) \\ &\quad + \max_{z^m} \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Bb}, \mathcal{F}, x^m \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Bb}, \mathcal{G}, z^m \right). \end{aligned}$$

Thus,

$$\begin{aligned} &\mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : \mathbb{E}[A_{f,g}] - \frac{1}{N} \sum_i k_i - \frac{1}{N} \sum_j p_j \geq \epsilon(\alpha + \beta + \mathbb{E}[A_{f,g}]) \right\} \\ &\leq \frac{32}{7} \max_{z^N} \mathcal{N}_1 \left(\frac{\epsilon\beta}{40B}, \mathcal{F}, x^N \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon\beta}{40(\lambda+1)B}, \mathcal{G}, z^N \right) \cdot \exp \left(-\frac{N\epsilon^2(1-\epsilon)\alpha}{512(\lambda+1)B^2} \right) \\ &\quad + \frac{64}{7} \max_{z^n} \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Ba}, \mathcal{F}, x^n \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640B(\lambda+1)a}, \mathcal{G}, z^n \right) \cdot \exp \left(-\frac{\epsilon^2(\alpha+\beta)N}{480B^2(\lambda+1)} \right) \\ &\quad + \frac{64}{7} \max_{z^m} \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Bb}, \mathcal{F}, x^m \right) \cdot \mathcal{N}_1 \left(\frac{\epsilon(\alpha+\beta)}{640Bb}, \mathcal{G}, z^m \right) \cdot \exp \left(-\frac{\epsilon^2(\alpha+\beta)N}{480B^2(\lambda+1)} \right) \\ &\leq 10\mathfrak{M}_{\mathcal{F}, \mathcal{G}}^{n,m,\lambda} \exp \left(-\frac{\epsilon^2(1-\epsilon)\alpha N}{512B^2(\lambda+1)} \right). \end{aligned}$$

This completes the proof of Theorem C.2. $\square$

C.3. Proof of Theorem 2.7

Write

$$D(f, g) := \mathcal{L}(f, g; \lambda) - \mathcal{L}(f^*, g^*; \lambda), \quad \widehat{D}(f, g) := \widehat{\mathcal{L}}(f, g; \lambda) - \widehat{\mathcal{L}}(f^*, g^*; \lambda).$$

In Theorem C.2, set $\epsilon = 1/2$, $\alpha = \beta = t/2$, and $C_0 = 1/(8192B^2)$. Since

$$D - \widehat{D} \geq \frac{1}{2}(t + D) \iff D - 2\widehat{D} \geq t,$$

that theorem gives

$$\begin{aligned} & \mathbb{P} \left\{ \exists (f, g) \in \mathcal{F} \times \mathcal{G} : D(f, g) - 2\widehat{D}(f, g) \geq t \right\} \\ & \leq 10(M_n(t) + M_m(t) + M_N(t)) \exp \left(-\frac{C_0 N t}{\lambda + 1} \right), \end{aligned}$$

where

$$\begin{aligned} M_n(t) &:= \max_{z^n} \mathcal{N}_1 \left(\frac{t}{1280Ba}, \mathcal{F}, x^n \right) \mathcal{N}_1 \left(\frac{t}{1280B(\lambda + 1)a}, \mathcal{G}, z^n \right), \\ M_m(t) &:= \max_{z^m} \mathcal{N}_1 \left(\frac{t}{1280Bb}, \mathcal{F}, x^m \right) \mathcal{N}_1 \left(\frac{t}{1280Bb}, \mathcal{G}, z^m \right), \\ M_N(t) &:= \max_{z^N} \mathcal{N}_1 \left(\frac{t}{160B}, \mathcal{F}, x^N \right) \mathcal{N}_1 \left(\frac{t}{160B(\lambda + 1)}, \mathcal{G}, z^N \right). \end{aligned}$$

Let

$$M := \sup_{z^N} \mathcal{N}_{1,\infty}(\epsilon_N, \mathcal{F} \times \mathcal{G}, z^N), \quad E_0 := \log M,$$

so that $\mathfrak{E}_{\mathcal{F} \times \mathcal{G}}^{N,\lambda} = 1 + E_0$. Restricting a full-sample ϵ_N -cover to the labeled and unlabeled blocks gives covers of radii ϵ_N/a and ϵ_N/b , respectively. For $t \geq 1/N$, the radii defining $M_n(t)$, $M_m(t)$, and $M_N(t)$ are at least the corresponding restricted-cover radii. Projecting a joint max-norm cover onto its two coordinates therefore gives

$$M_n(t) \leq M^2, \quad M_m(t) \leq M^2, \quad M_N(t) \leq M^2.$$

It follows that, for the data-dependent pair $(\hat{f}, \hat{g})$ and $t \geq 1/N$,

$$\mathbb{P} \left\{ D(\hat{f}, \hat{g}) - 2\widehat{D}(\hat{f}, \hat{g}) \geq t \right\} \leq 30M^2 \exp \left(-\frac{C_0 N t}{\lambda + 1} \right).$$

Set

$$\delta := \frac{\lambda + 1}{C_0 N} (\log 30 + 2E_0).$$

Because $B \geq 1$, we have $\delta \geq 1/N$. Integrating the preceding tail bound yields

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} \left[D(\hat{f}, \hat{g}) - 2\widehat{D}(\hat{f}, \hat{g}) \right] \\ & \leq \delta + 30M^2 \int_{\delta}^{\infty} \exp \left(-\frac{C_0 N t}{\lambda + 1} \right) dt \\ & = \frac{\lambda + 1}{C_0 N} (1 + \log 30 + 2E_0) \\ & \leq \frac{C_1 B^2 (\lambda + 1) \mathfrak{E}_{\mathcal{F} \times \mathcal{G}}^{N,\lambda}}{N}, \end{aligned}$$

for a universal constant $C_1 > 0$. Since $\widehat{D}(\hat{f}, \hat{g}) = \widehat{\Delta}(\hat{f}, \hat{g})$, we obtain

$$\begin{aligned} & \mathbb{E}_{\mathcal{D}} [\mathcal{L}(\hat{f}, \hat{g}; \lambda)] - \mathcal{L}(f^*, g^*; \lambda) \\ & \leq \frac{C_1 B^2 (\lambda + 1) \mathfrak{E}_{\mathcal{F} \times \mathcal{G}}^{N,\lambda}}{N} + 2\mathbb{E}_{\mathcal{D}} [\widehat{\Delta}(\hat{f}, \hat{g})]. \end{aligned} \tag{21}$$

Combining Corollary 2.3 and (21) completes the proof.

C.4. Proof of Remark 2.4 (ii)

Note that for any square-integrable random functions k_1 and k_2 , we have

$$1 - \frac{\mathbb{E}[\langle k_1, k_2 \rangle]^2]}{\mathbb{E}[\|k_1\|^2]\mathbb{E}[\|k_2\|^2]} = \frac{\mathbb{E}[\|k_1 - ck_2\|^2]}{\mathbb{E}[\|k_1\|^2]},$$

where

$$c = \frac{\mathbb{E}[\langle k_1, k_2 \rangle]}{\mathbb{E}[\|k_2\|^2]}.$$

Applying this identity with $k_1 = g^*(Z) - \hat{g}(Z)$ and $k_2 = f^*(X) - \hat{f}(X)$, we obtain

$$\begin{aligned} 1 - \rho_\star^2 &= \frac{\mathbb{E}_{\mathcal{D}, Z} \left[\left(g^*(Z) - \hat{g}(Z) - c(f^*(X) - \hat{f}(X)) \right)^2 \right]}{\mathbb{E}_{\mathcal{D}, Z} \left[\left(g^*(Z) - \hat{g}(Z) \right)^2 \right]} \\ &\geq \frac{\mathbb{E}_{\mathcal{D}, X} \left[\text{Var} \left(g^*(Z) - \hat{g}(Z) - c(f^*(X) - \hat{f}(X)) \mid \mathcal{D}, X \right) \right]}{\mathbb{E}_{\mathcal{D}, Z} \left[\left(g^*(Z) - \hat{g}(Z) \right)^2 \right]} \\ &= \frac{\mathbb{E}_{\mathcal{D}, X} \left[\text{Var} \left(g^*(Z) - \hat{g}(Z) \mid \mathcal{D}, X \right) \right]}{\mathbb{E}_{\mathcal{D}, Z} \left[\left(g^*(Z) - \hat{g}(Z) \right)^2 \right]}, \end{aligned}$$

where the inequality follows from the law of total variance.

Rearranging yields

$$\rho_\star^2 \leq \frac{\mathbb{E}_{\mathcal{D}, X} \left[\left(\mathbb{E} \left[g^*(Z) - \hat{g}(Z) \mid \mathcal{D}, X \right] \right)^2 \right]}{\mathbb{E}_{\mathcal{D}, Z} \left[\left(g^*(Z) - \hat{g}(Z) \right)^2 \right]}.$$

C.5. Proof of Theorem 3.1

For comparison element $h = (f, g)$ with atomic norms.

$$C_f := \|f\|_{L^1(\mathcal{D}_f)}, \quad C_g := \|g\|_{L^1(\mathcal{D}_g)}, \quad C_h := C_f + C_g.$$

Define the excess energy.

$$a_k := \|r_k\|_\star^2 - \|(Y, Y) - h\|_\star^2$$

and the norms

$$\alpha_k = \|\Pi_{S_k^f} r_k\|_\star, \quad \beta_k = \|\Pi_{S_k^g} r_k^{(g)}\|_\star.$$

By orthogonal projection properties,

$$\|r_{k+1}\|_\star^2 = \|r_k^{(g)}\|_\star^2 - \beta_k^2 = \|r_k\|_\star^2 - \alpha_k^2 - \beta_k^2 = \|r_k\|_\star^2 - c_k^2,$$

where $c_k^2 = \alpha_k^2 + \beta_k^2 = a_k - a_{k+1}$. To prove Theorem 3.1, we first prove the following lemmas.

Lemma C.4. For every $k \geq 1$,

$$A_k = \sup_{a \in \mathcal{D}_f^*} \langle r_k, a \rangle_\star \leq R_\lambda \alpha_k, \quad B_k = \sup_{b \in \mathcal{D}_g^*} \langle r_k, b \rangle_\star \leq R_\lambda \alpha_k + R_\lambda \beta_k.$$

Proof. Write $r_k = r_k^{(g)} + \Pi_{S_k^f} r_k$ with $r_k^{(g)} \perp S_k^f$ by Step 1. For any $b \in \mathcal{D}_g^*$,

$$\langle r_k, b \rangle_\star = \langle r_k^{(g)}, b \rangle_\star + \langle \Pi_{S_k^f} r_k, b \rangle_\star \leq \sup_{b \in \mathcal{D}_g^*} \langle r_k^{(g)}, b \rangle_\star + R_\lambda \|\Pi_{S_k^f} r_k\|_\star,$$

where the last inequality uses Cauchy–Schwarz and $\|b\|_\star \leq R_\lambda$. For $a \in \mathcal{D}_f^*$, by definition of ψ_k^* we have

$$\sup_{a \in \mathcal{D}_f^*} \langle r_k, a \rangle_\star \leq R_\lambda \sup_{a \in \mathcal{D}_f^*} \|\Pi_{S_{k-1}^f \cup \{a\}} r_k\|_\star = R_\lambda \|\Pi_{S_k^f} r_k\|_\star = R_\lambda \alpha_k.$$

Likewise,

$$\sup_{b \in \mathcal{D}_g^*} \langle r_k^{(g)}, b \rangle_* \leq R_\lambda \|\Pi_{S_k^g} r_k^{(g)}\|_* = R_\lambda \beta_k,$$

hence $\sup_b \langle r_k, b \rangle_* \leq R_\lambda \beta_k + R_\lambda \alpha_k$. This gives the second bound. $\square$

Lemma C.5. *We have*

$$\langle r_k, (Y, Y) \rangle_* = \|r_k\|_*^2 + \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_*, \quad (f_k, 0) = (f_{k-1}, 0) + \Pi_{S_k^f} r_k, \quad (22)$$

and, by polarization of the update,

$$\langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_* = \frac{1}{2} \left(\|(f_k, 0)\|_*^2 - \|(f_{k-1}, 0)\|_*^2 - \|\Pi_{S_k^f} r_k\|_*^2 \right). \quad (23)$$

Equivalently,

$$\langle r_k, (Y, Y) \rangle_* = \|r_k\|_*^2 + \frac{1}{2} \left(\|(f_k, 0)\|_*^2 - \|(f_{k-1}, 0)\|_*^2 - \|\Pi_{S_k^f} r_k\|_*^2 \right). \quad (24)$$

Proof. By definition $r_k = (Y, Y) - (f_{k-1}, g_{k-1})$, and the Step 1 update is the orthogonal projection in the $\star$ -inner product, hence

$$(f_k, g_{k-1}) = (f_{k-1}, g_{k-1}) + \Pi_{S_k^f} r_k, \quad \text{so} \quad (f_k, 0) = (f_{k-1}, 0) + \Pi_{S_k^f} r_k.$$

For (22), write

$$\langle r_k, (Y, Y) \rangle_* = \langle r_k, (f_{k-1}, g_{k-1}) + r_k \rangle_* = \|r_k\|_*^2 + \langle r_k, (f_{k-1}, g_{k-1}) \rangle_*.$$

Since $(f_{k-1}, 0) \in \text{span}(S_k^f)$ and $\Pi_{S_k^f} r_k$ is the orthogonal projection onto S_k^f ,

$$\langle r_k, (f_{k-1}, 0) \rangle_* = \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_*.$$

Moreover the previous step ends with $r_k \perp S_{k-1}^g$ and $(0, g_{k-1}) \in S_{k-1}^g$, hence $\langle r_k, (0, g_{k-1}) \rangle_* = 0$, proving (22).

For (23), expand the square using $(f_k, 0) = (f_{k-1}, 0) + \Pi_{S_k^f} r_k$.

$$\|(f_k, 0)\|_*^2 = \|(f_{k-1}, 0)\|_*^2 + 2\langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_* + \|\Pi_{S_k^f} r_k\|_*^2,$$

and rearrange to obtain (23). Substituting (23) into (22) yields (24). $\square$

Lemma C.6. *For every $k \geq 1$,*

$$\sum_{i=2}^{k+1} a_i \leq 2\sqrt{2}C_h R_\lambda \sum_{i=1}^k c_i, \quad C_h = C_f + C_g, \quad c_i = \sqrt{a_i - a_{i+1}}.$$

Proof. Let $e := (Y, Y) - h$, $r_k := (Y, Y) - (f_{k-1}, g_{k-1})$, and recall

$$a_k := \|r_k\|_*^2 - \|e\|_*^2, \quad c_k^2 = a_k - a_{k+1} = \alpha_k^2 + \beta_k^2.$$

By Lemma C.5,

$$\|r_k\|_*^2 = \langle r_k, (Y, Y) \rangle_* - \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_*.$$

Then, by arithmetic–geometric mean inequality

$$\begin{aligned} a_k &= \|r_k\|_*^2 - \|e\|_*^2 = \langle r_k, e \rangle_* + \langle r_k, (f, 0) \rangle_* + \langle r_k, (0, g) \rangle_* - \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_* - \|e\|_*^2 \\ &\leq \frac{1}{2} a_k + \langle r_k, (f, 0) \rangle_* - \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_* + \langle r_k, (0, g) \rangle_*, \end{aligned}$$

and we use the bounds $\langle r_k, (f, 0) \rangle_* \leq C_f A_k$ and $\langle r_k, (0, g) \rangle_* \leq C_g B_k$ to control the two correlation terms, hence

$$\frac{a_k}{2} \leq C_f A_k + C_g B_k - \langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_*.$$

Apply the polarization identity (Lemma C.5) to the f -history:

$$\langle \Pi_{S_k^f} r_k, (f_{k-1}, 0) \rangle_\star = \frac{1}{2} (\|(f_k, 0)\|_\star^2 - \|(f_{k-1}, 0)\|_\star^2 - \alpha_k^2).$$

Therefore,

$$\frac{a_k}{2} \leq C_f A_k + C_g B_k + \frac{1}{2} (\|(f_{k-1}, 0)\|_\star^2 - \|(f_k, 0)\|_\star^2) + \frac{1}{2} \alpha_k^2.$$

Summing over $i = 1, \dots, k$ (with $f_0 = 0$) yields

$$\sum_{i=1}^k \frac{a_i}{2} \leq C_f \sum_{i=1}^k A_i + C_g \sum_{i=1}^k B_i - \frac{1}{2} \|(f_k, 0)\|_\star^2 + \frac{1}{2} \sum_{i=1}^k \alpha_i^2. \quad (25)$$

Since $\sum_{i=1}^k \alpha_i^2 \leq \sum_{i=1}^k c_i^2 = \sum_{i=1}^k (a_i - a_{i+1}) = a_1 - a_{k+1}$, $A_i \leq R_\lambda \alpha_i$, and $B_i \leq R_\lambda \alpha_i + R_\lambda \beta_i \leq \sqrt{2} R_\lambda c_i$, we get

$$C_f \sum_{i=1}^k A_i + C_g \sum_{i=1}^k B_i \leq C_f R_\lambda \sum_{i=1}^k c_i + \sqrt{2} C_g R_\lambda \sum_{i=1}^k c_i \leq \sqrt{2} (C_f + C_g) R_\lambda \sum_{i=1}^k c_i = \sqrt{2} C_h R_\lambda \sum_{i=1}^k c_i.$$

Dropping the nonpositive term $-\frac{1}{2} \|(f_k, 0)\|_\star^2$ in (25) gives

$$\sum_{i=1}^k \frac{a_i}{2} \leq \sqrt{2} C_h R_\lambda \sum_{i=1}^k c_i + \frac{a_1 - a_{k+1}}{2},$$

which implies

$$\sum_{i=2}^{k+1} a_i \leq 2\sqrt{2} C_h R_\lambda \sum_{i=1}^k c_i.$$

□

We are ready to prove Theorem 3.1.

Proof. Set

$$S_k := \sum_{i=1}^k a_i, \quad H_k := \sum_{i=1}^k \frac{1}{i}.$$

We have that for every $k \geq 1$,

$$\sum_{i=2}^{k+1} a_i \leq 2\sqrt{2} C_h R_\lambda \sum_{i=1}^k c_i, \quad c_i = \sqrt{a_i - a_{i+1}}. \quad (26)$$

Summation by parts (discrete Abel transform). For every $k \geq 1$,

$$\sum_{i=1}^k i(a_i - a_{i+1}) = \sum_{i=1}^k i c_i^2 = S_k - k a_{k+1}. \quad (27)$$

Indeed,

$$\sum_{i=1}^k i(a_i - a_{i+1}) = \sum_{i=1}^k \sum_{j=1}^i (a_i - a_{i+1}) = \sum_{j=1}^k \sum_{i=j}^k (a_i - a_{i+1}) = \sum_{j=1}^k (a_j - a_{k+1}) = S_k - k a_{k+1}.$$

Cauchy–Schwarz with weights and (27). By Cauchy–Schwarz,

$$\left(\sum_{i=1}^k c_i \right)^2 = \left(\sum_{i=1}^k \frac{1}{\sqrt{i}} \sqrt{i} c_i \right)^2 \leq \left(\sum_{i=1}^k \frac{1}{i} \right) \left(\sum_{i=1}^k i c_i^2 \right) = H_k (S_k - k a_{k+1}) \leq H_k S_k.$$

Using (26) and $S_k \leq S_{k+1}$,

$$S_{k+1} - a_1 = \sum_{i=2}^{k+1} a_i \leq 2\sqrt{2}C_h R_\lambda \sum_{i=1}^k c_i \leq 2\sqrt{2}C_h R_\lambda \sqrt{H_k S_{k+1}}.$$

Then

$$S_{k+1} \leq a_1 + 2\sqrt{2}C_h R_\lambda \sqrt{H_k S_{k+1}} \leq a_1 + S_{k+1}/2 + 4C_h^2 R_\lambda^2 H_k,$$

so $S_{k+1} \leq 2a_1 + 8C_h^2 R_\lambda^2 H_k$. Hence for all $k \geq 1$,

$$S_k \leq 2a_1 + 8C_h^2 R_\lambda^2 H_k. \quad (28)$$

Pointwise rate. Since (a_k) is decreasing, $ka_k \leq S_k$. Combining with (28),

$$a_k \leq \frac{2a_1}{k} + \frac{8C_h^2 R_\lambda^2 H_k}{k} \leq C \frac{C_h^2 R_\lambda^2 \log(k+1)}{k},$$

for a universal constant $C > 0$, completing the proof. $\square$