Journal of Machine Learning Research 25 (2024) 1-37 Submitted 10/22; Revised 7/23; Published 1/24

Monotonic Risk Relationships under Distribution Shifts for Regularized Risk Minimization

Daniel LeJeune DANIEL@DLEJ.NET
Department of Statistics
Stanford University
Stanford, CA 94305-4020, USA

Jiayu Liu JIAYU.LIU@TUM.DE
Reinhard Heckel REINHARD.HECKEL@TUM.DE
Department of Electrical and Computer Engineering
Technical University of Munich
80333 Munich, DE

Editor: Daniel Hsu

Abstract

Machine learning systems are often applied to data that is drawn from a different distribution than the training distribution. Recent work has shown that for a variety of classification and signal reconstruction problems, the out-of-distribution performance is strongly linearly correlated with the in-distribution performance. If this relationship or more generally a monotonic one holds, it has important consequences. For example, it allows to optimize performance on one distribution as a proxy for performance on the other. In this paper, we study conditions under which a monotonic relationship between the performances of a model on two distributions is expected. We prove an exact asymptotic linear relation for squared error and a monotonic relation for misclassification error for ridge-regularized general linear models under covariate shift, as well as an approximate linear relation for linear inverse problems.

Keywords: distribution shifts, asymptotics, empirical risk minimization, general linear models, inverse problems

1. Introduction

Machine learning models are typically evaluated by shuffling a set of labeled data, splitting it into training and test sets, and evaluating the model trained on the training set on the test set. This measures how well the model performs on the distribution the model was trained on. However, in practice a model is most commonly not applied to such in-distribution data, but rather to out-of-distribution data that is almost always at least slightly different. In order to understand the performance of machine learning methods in practice, it is therefore important to understand how out-of-distribution performance relates to in-distribution performance.

While there are settings in which models with similar in-distribution performance have different out-of-distribution performance (McCoy et al., 2020), a series of recent empirical studies have shown that often, the in-distribution and out-of-distribution performances of models are strongly correlated:

- Recht et al. (2019), Yadav and Bottou (2019), and Miller et al. (2020) constructed new test sets for the popular CIFAR-10, ImageNet, and MNIST image classification problems and for the SQuAD question answering datasets by following the original data collection and labeling process as closely as possible. For CIFAR-10 and ImageNet the performance drops significantly when evaluated on the new test set, indicating that even when following the original data collection and labeling process, a significant distribution shift can occur. In addition, for all four distribution shifts, the in- and out-of-distribution errors are strongly linearly correlated.

- Miller et al. (2021) identified a strong linear correlation of the performance of image classifiers for a variety of natural distribution shifts. Apart from classification, the linear performance relationship phenomenon is also observed in machine learning tasks where models produce real-valued output, for example in pose estimation (Miller et al., 2021) and object detection (Caine et al., 2021).

- Darestani et al. (2021) identified a strong linear correlation of the performance of image reconstruction methods for a variety of natural distribution shifts. This relation between in- and out-of-distribution performances persisted for image reconstruction methods that are only tuned, i.e., only a small set of hyperparameters is chosen based on hyperparameter optimization on the training data.

An important consequence of a linear, or more generally, a monotonic relationship between in- and out-of-distribution performances is that a model that performs better in-distribution also performs better on out-of-distribution data, and thus measuring in-distribution performance can serve as a proxy for tuning and comparing different models for application on out-of-distribution data.

It is therefore important to understand when a linear or more generally a monotonic relationship between the performance on two distributions occurs. In this paper we study this question theoretically and empirically for a class of distribution shifts where the feature or signal models come from different distributions, also known as covariate shift.

First, we show that for a real-world regression problem, in- and out-of-distribution performances are linearly correlated. Specifically, we show that for object detection, the performance of models trained on the COCO 2017 training set and evaluated on the COCO 2017 validation set is linearly correlated with the performance on the VOC 2012 dataset. This finding establishes that a linear risk relation also occurs for regression problems, beyond classification problems as established before.

We then consider a simple linear regression model with a feature vector drawn from a different subspace for in- and out-of-distribution data. We provide sufficient conditions for a linear estimator that characterizes when a linear relation between in- and out-of-distribution occurs.

Next, we consider a general setup encompassing classification and regression, and consider a distribution shift model on the feature vectors. We consider a large class of estimators obtained with regularized empirical risk minimization, and show that as various training parameters change, including for example the regularization strength or the number of training examples (resulting in different estimators), the relationship between in- and out-of-distribution performances is monotonic. Different classes of estimators follow different monotonic relations, and we also observe this in practice (see Figure 3). Interestingly, for a certain class of shifts in classification, we recover a linear relation for a nonlinear function of the risks that is remarkably similar to that demonstrated empirically by Miller et al. (2021).

Finally, we study linear inverse problems, to understand when a linear relation occurs in a signal reconstruction problem. We consider a distribution shift model consisting of a shift in subspace as well as noise variance, and again characterize conditions under which a linear or near-linear relation between in- and out-of-distribution performances exists.

Our results suggest that linear risk relationships observed in regression and classification actually arise by independent mechanisms, being based on a shift in feature subspace for regression and a shift in feature scaling for classification.

Code for the experiments and figures in this paper can be found at `https://github.com/MLI-lab/monotonic_risk_relationships`.

1.1 Prior Theoretical Work on Characterizing Linear Performance Relations

Classical theory for characterizing out-of-distribution performance ensures that the difference between in- and out-of-distribution performance of an estimator is bounded by a function of the distance of the training and test distributions (Quiñonero-Candela et al., 2008; Ben-David et al., 2010; Cortes and Mohri, 2014). Such bounds often apply to a class of target distributions. In contrast, we are interested in precise relationships between a fixed source and target distribution.

Regarding characterizing linear relationships, Miller et al. (2021, Sec. 7) proved that for a distribution shift for a binary mixture model, the in- and out-of-distribution accuracies have a linear relation if the features vectors are sufficiently high-dimensional. Mania and Sra (2020) showed that an approximate linear relationship occurs under a model similarity assumption that high accuracy models correctly classify most of the data points that are correctly classified by lower accuracy models.

Most related to our work is that of Tripuraneni et al. (2021), who revealed an exact linear relation for squared error of a linear random feature regression model under a covariate shift in the high-dimensional limit. This covariate shift is philosophically similar to the simplifying assumption we make for the main statement and interpretation of our results, and yields a similar linear relation for squared error. However, our results apply to a broader class of general linear models and extend to misclassification error, and we go further to capture how the distribution shift can depend on the task itself, which captures how classification problems can become easier or harder. Moreover, our results predict general monotonic relationships as opposed to only linear ones.

2. Linear Relations in Regression and Motivation for the Subspace Model

Prior work in the distribution shift literature for prediction tasks has focused on either classification or on problems with real-valued outputs but using discrete performance metrics—for example, pose estimation (Miller et al., 2021) and object detection (Caine et al., 2021). Here, we consider a real-valued squared error metric and show that linear relationships between in- and out-of-distribution performances also occur in a standard regression setup.

We evaluate a collection of neural network models for object detection, which are trained on the COCO 2017 training set (Lin et al., 2014): Faster R-CNN (Ren et al., 2015), Mask

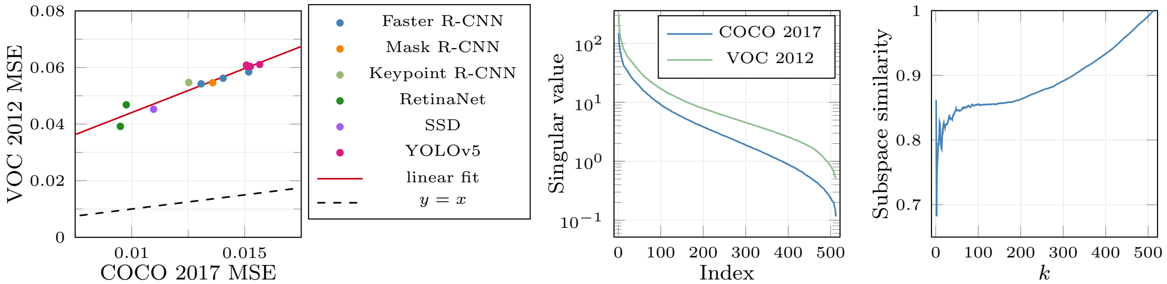


Figure 1: Bounding box prediction on COCO 2017 and VOC 2012 datasets. **Left:** There is an approximate linear relation of mean squared error (MSE) for models trained COCO 2017. **Middle:** The spectrum of the feature spaces of YOLOv5 on the two datasets decays quickly, which suggests that a feature subspace model could be a reasonable approximation. **Right:** A principal-angle-based similarity between subspaces spanned by the top k principal components on the two datasets. The subspaces are well-aligned, which is a sufficient condition for a linear relation as stated in Theorem 1.

R-CNN (He et al., 2017), Keypoint R-CNN (He et al., 2017), SSD (Liu et al., 2016), RetinaNet (Lin et al., 2017), and YOLOv5 (Redmon et al., 2016; Jocher et al., 2020). See Figure 1 (left), where we compute their mean squared errors for bounding box coordinate prediction on the COCO 2017 validation set and the VOC 2012 training/validation set (Everingham et al., 2010). The models we consider all perform worse on the out-of-distribution data, and the in- and out-of-distribution performances are approximately linearly related.

It is in general difficult to model distribution shifts analytically. In this work, one aspect of distribution shifts that we model is the change in the subspaces where the feature vectors lie. To motivate this model, we next examine the feature space of the YOLOv5 model on the in- and out-of-distribution data.

The YOLOv5 model, and all other models considered, can be viewed to make a prediction for an image by generating features through several layers, and then aggregating those features with a linear layer (or a very shallow neural network) to make a prediction. We consider the 512-dimensional feature vectors from the penultimate layer of YOLOv5 as the features. Let $\{\mathbf{z}_j^{(i)} \in \mathbb{R}^{512} : i \in [N_{\text{in}}], j \in [K_{\text{in}}^{(i)}]\}$ and $\{\mathbf{z}_j^{(i)} \in \mathbb{R}^{512} : i \in [N_{\text{out}}], j \in [K_{\text{out}}^{(i)}]\}$ be the set of feature vectors of the in- and out-of-distribution data, respectively, where $\mathbf{z}_j^{(i)}$ is the feature vector of the j^{th} true positive prediction on image i, N_{in} and N_{out} are the numbers of images of the respective datasets, and $K_{\text{in}}^{(i)}$ and $K_{\text{out}}^{(i)}$ are the number of true positive predictions on the i^{th} images of the respective datasets. We perform principal component analysis on these two sets of feature vectors and plot the spectra in Figure 1 (middle). We observe that approximating the feature space by the top 100 principal components explains 96.0% and 95.6% of the variances of COCO 2017 and VOC 2012 respectively. This observation demonstrates that the feature vectors approximately lie in subspaces of the full feature space.

Moreover, Figure 1 (right) shows that the feature subspaces for the two distributions are overlapping substantially. Specifically, Figure 1 (right) shows the subspace similarity defined as $\sqrt{\|\cos(\boldsymbol{\theta})\|_2^2/k}$ (Soltanolkotabi and Candès, 2012; Heckel and Bölcskei, 2015), where $\boldsymbol{\theta}$ is the vector of principal angles between the subspaces spanned by the top k

principal components of the individual feature vector sets. The subspaces spanned by the top 100 principal components, which account for over 95% of the variance, have a 0.855 subspace similarity (note that the maximum value 1 is achieved for $\boldsymbol{\theta} = \mathbf{0}$). More details on the experiment are in Appendix A.1.

Because the output of neural networks is simply a linear model applied to this feature space, this observation suggests that the relationship between in- and out-of-distribution performances of even highly nonlinear models such as neural networks on data from highly nonlinear spaces may be modeled by a change in linear subspaces of a transformed feature space. Therefore, we theoretically study the effect of changes of subspace in linear models and the resulting performance relationships. Our results consider fixed feature spaces, while different deep learning models have different feature representations at the final layer. However, our study can shed light on performance changes of models from the same family that share similar feature representations under distribution shifts.

3. Linear Relations in Regression in Finite Dimensions

We begin our theoretical study by considering the linear regression setting under additive noise: $y = \mathbf{x}^\mathsf{T}\boldsymbol{\beta}^* + z$, where $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is a fixed parameter vector that determines the model, and z is independent observation noise. We assume that the feature vector $\mathbf{x}$ is drawn randomly from a subspace, also known as the hidden manifold model (Goldt et al., 2020). Let $d_P, d_Q \leq d$. For data from distribution P, the feature vector is given by $\mathbf{x} = \mathbf{U}_P\mathbf{c}_P$, where $\mathbf{U}_P \in \mathbb{R}^{d \times d_P}$ has orthonormal columns and $\mathbf{c}_P \in \mathbb{R}^{d_P}$ is zero-mean and has identity covariance. The noise variable is zero-mean and has variance σ_P^2. The data from distribution Q is generated in the same manner, but the signal is from a different subspace with $\mathbf{x} = \mathbf{U}_Q\mathbf{c}_Q$, where $\mathbf{U}_Q \in \mathbb{R}^{d \times d_Q}$ has orthonormal columns, $\mathbf{c}_Q \in \mathbb{R}^{d_Q}$ is zero-mean and has identity covariance, and the noise is zero-mean and has variance σ_Q^2.

For an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$, define the risk on distribution P with respect to the squared error metric as $\mathcal{R}_P(\widehat{\boldsymbol{\beta}}) = \mathbb{E}_{\mathbf{x}\sim P}\left[(y - \mathbf{x}^\top\widehat{\boldsymbol{\beta}})^2\right]$ (respectively $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}})$ on distribution Q). We are interested in the relation of those risks for a class of estimators. We consider an estimate of the model parameter $\boldsymbol{\beta}^*$ assuming knowledge of the distribution for simplicity, equivalent to having large amounts of training data. The analysis can be extended readily to estimates based on finite samples. We consider the estimator

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \mathbb{E}_P\left[(\boldsymbol{\beta}^\mathsf{T}\mathbf{x} - y)^2\right] + \lambda\|\boldsymbol{\beta}\|_2^2,$$

parameterized by the regularization parameter λ. It can be shown that $\widehat{\boldsymbol{\beta}}_\lambda = \alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\beta}^*$ for $\alpha = 1/(1+\lambda)$, which is the projection of $\boldsymbol{\beta}^*$ onto the subspace scaled by the factor $\alpha \in [0, 1]$.

The following theorem provides sufficient conditions for a linear relation between the in- and out-of-distribution risks $\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)$ of this class of estimators parameterized by the regularization parameter λ. Theorem 1 is a consequence of Theorem 6 in Appendix C, which also provides a necessary condition for a linear risk relation.

Theorem 1 (Sufficient conditions) *The out-of-distrubiton risk $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)$ is an affine function of the in-distribution risk $\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)$ as a function of the regularization parameter λ if one of the following conditions holds:*

(a) range$(\mathbf{U}_Q) \subseteq$ range$(\mathbf{U}_P)$, or range$(\mathbf{U}_P) \subseteq$ range$(\mathbf{U}_Q)$;

(b) $\boldsymbol{\beta}^* \in$ range$(\mathbf{U}_P)$.

Moreover, for random $\boldsymbol{\beta}^*$, the expected out-of-distribution risk, $\mathbb{E}_{\boldsymbol{\beta}^*}\left[\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)\right]$, is an affine function of the expected in-distribution risk $\mathbb{E}_{\boldsymbol{\beta}^*}\left[\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)\right]$ if

(c) $\mathbb{E}\left[\boldsymbol{\beta}^*\boldsymbol{\beta}^{*\top}\right] = \mathbf{I}$.

Condition (a) is a property of the distribution shift itself. When the subspaces are aligned between the two distributions, we observe a linear risk relationship for the set of estimators parameterized by λ. Recall from the previous section, that the feature subspaces of the object detection model we evaluate roughly align, as shown in Figure 1 (right). Thus, our theorem suggests a linear relationship, which in turn sheds light on the linear relationship we observed in practice. We remark that the linear relationship guaranteed by Theorem 1 is exact assuming full knowledge of the source distribution, but only approximate in the finite sample regime for an estimate that minimizes the regularized empirical risk.

Condition (b) is a property of the parameter vector $\boldsymbol{\beta}^*$ and its learnability under distribution P. Under condition (b), $\widehat{\boldsymbol{\beta}}_\lambda = \alpha\boldsymbol{\beta}^*$, which greatly simplifies the risks:

$$\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda) = (1-\alpha)^2\boldsymbol{\beta}^{*\top}\boldsymbol{\beta}^* + \sigma_P^2 \quad \text{and} \quad \mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda) = (1-\alpha)^2\boldsymbol{\beta}^{*\top}\mathbf{U}_Q\mathbf{U}_Q^\top\boldsymbol{\beta}^* + \sigma_Q^2.$$

It is thus very clear that there is a monotonic relation, as both are affine in $(1-\alpha)^2$.

Condition (c) meanwhile is a property of randomness in $\boldsymbol{\beta}^*$ that leads to the elimination of interaction terms that would prevent a monotonic relation. While the above result is given for the expectation, the same effect would also occur for single problem instances in high dimensions due to concentration of measure.

The intuition behind these three conditions all carry over to our more general results.

4. Asymptotic Monotonic Relations for General Linear Models

In the previous section, we demonstrated a linear risk relationship under a subspace shift for linear regression models. In this section, we provide a much more general result that holds for a larger class of distribution shifts, setups (i.e., regression and classification), and estimators, specifically for a class of estimators based on regularized empirical risk minimization.

4.1 Linear Model Framework

We consider a general framework of linear models $f(\mathbf{x}) = \phi(\mathbf{x}^\top\boldsymbol{\beta})$ for some $\boldsymbol{\beta} \in \mathbb{R}^d$, labeling function $\phi\colon \mathbb{R} \to \mathbb{R}$, and centralized Gaussian data under two distributions

$$P\colon \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \tfrac{1}{d}\boldsymbol{\Sigma}_P) \quad \text{and} \quad Q\colon \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \tfrac{1}{d}\boldsymbol{\Sigma}_Q),$$

where $\mathbf{\Sigma}_P$ and $\mathbf{\Sigma}_Q$ are positive semidefinite covariance matrices. Given a ground truth model $f^*(\mathbf{x}) = \phi(\mathbf{x}^\top \boldsymbol{\beta}^*)$, we define the *risk* of a model $f(\mathbf{x}) = \phi(\mathbf{x}^\top \boldsymbol{\beta})$ with respect to an error metric $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ on distributions P and Q as

$$\mathcal{R}_P(\boldsymbol{\beta}) \triangleq \mathop{\mathbb{E}}_{\mathbf{x}\sim P}\left[\psi(\mathbf{x}^\top\boldsymbol{\beta}^*, \mathbf{x}^\top\boldsymbol{\beta})\right] \quad \text{and} \quad \mathcal{R}_Q(\boldsymbol{\beta}) \triangleq \mathop{\mathbb{E}}_{\mathbf{x}\sim Q}\left[\psi(\mathbf{x}^\top\boldsymbol{\beta}^*, \mathbf{x}^\top\boldsymbol{\beta})\right].$$

We consider the squared error $\psi(z^*, z) = (z^* - z)^2$ and misclassification error $\psi(z^*, z) = \mathbb{1}\{z^* z < 0\}$ as error metrics for regression and classification, respectively. Now define the random variables, often referred to as the decision functions,

$$(Z_P^*, Z_P) = (\mathbf{x}^\top\boldsymbol{\beta}^*, \mathbf{x}^\top\boldsymbol{\beta}) : \mathbf{x} \sim P \quad \text{and} \quad (Z_Q^*, Z_Q) = (\mathbf{x}^\top\boldsymbol{\beta}^*, \mathbf{x}^\top\boldsymbol{\beta}) : \mathbf{x} \sim Q.$$

As we capture in the following proposition, the risks for any linear model $f(\mathbf{x}) = \phi(\mathbf{x}^\top\boldsymbol{\beta})$ depend only on a few parameters defining the covariances of the decision functions.

Proposition 2 *The vectors (Z_P^*, Z_P) and (Z_Q^*, Z_Q) are zero-mean bivariate normal random vectors. Furthermore, $\mathcal{R}_P(\boldsymbol{\beta})$ and $\mathcal{R}_Q(\boldsymbol{\beta})$ are functions only of the covariance matrices $\operatorname{Cov}(Z_P^*, Z_P) \in \mathbb{R}^{2\times 2}$ and $\operatorname{Cov}(Z_Q^*, Z_Q) \in \mathbb{R}^{2\times 2}$, respectively.*

Thus, while the covariance matrices $\mathbf{\Sigma}_P$, $\mathbf{\Sigma}_Q$, and the model parameters $\boldsymbol{\beta}^*$ and $\boldsymbol{\beta}$ in general comprise on the order of d^2 parameters, the risks $\mathcal{R}_P(\boldsymbol{\beta})$ and $\mathcal{R}_Q(\boldsymbol{\beta})$ are characterized by no more than 6 parameters of the covariance matrices $\operatorname{Cov}(Z_P^*, Z_P)$ and $\operatorname{Cov}(Z_Q^*, Z_Q)$. In order to have a monotonic relation between the risks $\mathcal{R}_P(\boldsymbol{\beta})$ and $\mathcal{R}_Q(\boldsymbol{\beta})$ the dependency needs to be reduced to a single parameter, which requires additional assumptions on the class of models and the distribution shifts, which we state in the next subsection.

4.2 Asymptotic Estimation with Regularized Empirical Risk Minimization

We consider predictors $\hat{f} = \phi(\mathbf{x}^\top\widehat{\boldsymbol{\beta}})$ where the parameter $\widehat{\boldsymbol{\beta}}$ is the ridge-regularized empirical risk minimization (ERM) estimate

$$\widehat{\boldsymbol{\beta}}(\mathcal{D}, \ell, \lambda) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i^\top\boldsymbol{\beta}) + \frac{\lambda}{2}\|\boldsymbol{\beta}\|_2^2, \tag{1}$$

where $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is a training set with covariates $\mathbf{x}_i \sim P$, $\ell\colon \mathbb{R}^2 \to \mathbb{R}$ a loss function, and $\lambda > 0$ a regularization parameter.

In finite dimensions, determining the in- and out-of distribution risk via determining the covariances $\operatorname{Cov}(Z_P^*, Z_P)$ and $\operatorname{Cov}(Z_Q^*, Z_Q)$ even for linear models with convex loss functions is not possible in general, making the task of identifying a monotonic risk relation difficult. Fortunately, however, it has recently been shown (Thrampoulidis et al., 2018; Emami et al., 2020; Loureiro et al., 2021) that as the problem dimensionality becomes large, thanks to concentration of measure effects, the solution to regularized ERM problems can be characterized by the solution of a system of scalar fixed point equations in only a few variables. Our result relies on such an asymptotic characterization by Loureiro et al. (2021).

In the following, we state the asymptotic setup, data generation process, and distribution-shift model that we consider as an assumption, so that we can refer to it later.

Assumption A (Setup)

(A1) **Asymptotically proportional regime**. *The training data set size n and dimensionality d tend to infinity with fixed finite ratio d/n.*

(A2) **Training data generation.** *The training data $\mathcal{D}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ is from distribution P, and independently generated as $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{\Sigma}_P)$ and $y_i = \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}^*, \xi_i)$ for a labeling function $\varphi \colon \mathbb{R}^2 \to \mathbb{R}$, random noise ξ_i independent of $\mathbf{x}_i$ and ground truth coefficient vector $\boldsymbol{\beta}^*$. Additionally, $\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\|\mathbf{y}\|_2^2\right] < \infty$.*

(A3) **Ground-truth coefficient vector and structure of the covariances.** *The ground truth coefficient vector $\boldsymbol{\beta}^*$ has elements drawn i.i.d. from a zero-mean sub-Gaussian distribution with variance σ_β^2 and is independent of $\mathcal{D}_n$, and $\mathbf{\Sigma}_P = \mathbf{\Pi}_P$ is a projection operator onto a subspace of dimension d_P such that $d_P/d \to r_P \in (0, 1]$. Furthermore, the covariances $\mathbf{\Sigma}_P$ and $\mathbf{\Sigma}_Q$ are simultaneously diagonalizable.*

(A4) **Loss function of ERM.** *The loss function ℓ is a proper, lower semi-continuous, convex function that is pseudo-Lipschitz of order 2 (see Definition 7 in Appendix D for a formal definition) such that for all n and $c > 0$, if $\|\mathbf{z}\|_2 \le c\sqrt{n}$ then there exists a positive constant C such that $\sup_{\mathbf{z}' \in \partial_{\mathbf{z}} \bar{\ell}(\mathbf{y},\mathbf{z})} \|\mathbf{z}'\|_2 \le C\sqrt{n}$, where $\bar{\ell}(\mathbf{y}, \mathbf{z}) = \sum_{i=1}^n \ell(y_i, z_i)$. Furthermore, for the standard normal random vector $\mathbf{g} \in \mathbb{R}^n$, $\frac{1}{n}\mathbb{E}\left[\bar{\ell}(\mathbf{y}, \mathbf{g})\right]$ is uniformly bounded in n.*

The data generating process (A2) and the assumption on the loss function of ERM (A4) are standard for most convex and linear ERM formulations used in machine learning for regression and classification.

The assumptions on the ground truth coefficients and covariance matrices in Assumption A3 are stronger than necessary; our result can in fact even be proved for deterministic $\boldsymbol{\beta}^*$ and essentially arbitrary $\mathbf{\Sigma}_Q$ and non-isotropic $\mathbf{\Sigma}_P$ (see Appendix D). However, these assumptions greatly simplify the form of the results at little expense of generality.

Assumption A1 puts us in the proportional asymptotics regime, but the concentration effects are often realized at only modest data sizes; see Figure 2.

Under Assumption A, the ERM estimator has the form $\widehat{\boldsymbol{\beta}} = \mathbf{\Pi}_P(a\boldsymbol{\beta}^* + c\mathbf{g})$ for some $\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ independent of $\boldsymbol{\beta}^*$ (see Corollary 10), extending the intuition from Theorem 1. Therefore, the covariances $\mathrm{Cov}(Z_P^*, Z_P)$ and $\mathrm{Cov}(Z_Q^*, Z_Q)$ have only two degrees of freedom (a and c) in the asymptotic limit for fixed P, Q, and $\boldsymbol{\beta}^*$, even as we vary a number of different learning problem parameters such as loss function, noise level, labeling function, regularization strength, and number of training examples (see Lemma 11). Even with only two degrees of freedom, this is still not enough to imply a monotonic relation for general risks (see Section 4.4); however, remarkably, it turns out that this specific structure is sufficient for monotonicity for both squared error and misclassification error.

For this Setup (A), the monotonic relations between in- and out-of-distribution risks for distributions P and Q and ground truth $\boldsymbol{\beta}^*$ are entirely described by only three limiting scalar parameters of the distribution shift, which we define next. Our assumption that these quantities converge is stronger than necessary; we only need that these quantities be almost surely uniformly bounded (e.g., by making $\mathbf{\Sigma}_Q$ uniformly bounded in operator norm and have non-vanishing subspace overlap with $\mathbf{\Pi}_P$), in which case we can simply apply this assumption and our results to each convergence subsequence. However, to keep the statements of our results clean, we assume convergence of these parameters.

Assumption B (Parameters) *The following limits exist for $\boldsymbol{\beta}_P^* \triangleq \mathbf{\Pi}_P \boldsymbol{\beta}^*$ almost surely:*

$$\gamma = \lim_{d\to\infty} \frac{\boldsymbol{\beta}_P^{*\top} \mathbf{\Sigma}_Q \boldsymbol{\beta}_P^*}{d_P \sigma_\beta^2}, \qquad \mu = \lim_{d\to\infty} \frac{\boldsymbol{\beta}^{*\top} \mathbf{\Sigma}_Q \boldsymbol{\beta}^*}{\boldsymbol{\beta}_P^{*\top} \mathbf{\Sigma}_Q \boldsymbol{\beta}_P^*} \geq 1, \qquad \kappa = \lim_{d\to\infty} \frac{\mathrm{tr}[\mathbf{\Sigma}_Q \mathbf{\Pi}_P]}{d_P}.$$

The parameters γ and μ are straightforward to interpret. The parameter γ captures the ratio of the energy of $\boldsymbol{\beta}_P^*$ as measured by $\mathbf{\Sigma}_Q$ compared to $\mathbf{\Sigma}_P$. If $\mathbf{\Sigma}_Q$ is a scaled projection operator $\tau\mathbf{\Pi}_Q$ for some $\tau > 0$ with d_{PQ} dimensions overlapping with $\mathbf{\Pi}_P$, then $\gamma = \tau d_{PQ}/d_P$. The parameter μ captures the ratio of the total energy of $\boldsymbol{\beta}^*$ as measured by $\mathbf{\Sigma}_Q$ compared to its restriction to the subspace determined by $\mathbf{\Pi}_P$. For the same scaled projection operator example, if d_Q is the dimension of the subspace of $\mathbf{\Pi}_Q$, then $\mu = d_Q/d_{PQ}$.

The parameter κ introduces the nuance of *task dependence* of the distribution shift. Note that the ground-truth parameter $\boldsymbol{\beta}^*$ and the covariance matrix $\mathbf{\Sigma}_Q$ might be statistically correlated. (We might like to consider $\mathbf{\Sigma}_Q$ to be deterministic, whereas $\boldsymbol{\beta}^*$ is a random variable. However, our results in Appendix D hold almost surely for a fixed, deterministic, covariance–ground-truth pair $(\mathbf{\Sigma}_Q, \boldsymbol{\beta}^*)$; so we can think about this pair as deterministic or correlated). As an example of such a correlation, $\mathbf{\Sigma}_Q$ may have larger eigenvalues in the directions where $\boldsymbol{\beta}^*$ is larger in magnitude, and therefore $\gamma > \kappa$. Intuitively, since we assume $\mathbf{\Sigma}_P$ to be isotropic on the subspace, this means that at test time, the prediction depends more on coefficients that were learned better during training, making the problem easier. Conversely, if $\gamma < \kappa$, $\mathbf{\Sigma}_Q$ and $\boldsymbol{\beta}^*$ are anti-correlated, and the prediction becomes more difficult since features that were learned poorly are emphasized more highly. This can be summarized with the ratio κ/γ, which when less than 1 implies an easier distribution shift, and when greater than 1 implies a harder one. When $\gamma = \kappa$, we say the shift is *task-independent*. The case of *task-dependent* shifts where $\kappa \neq \gamma$ cannot be captured by the $\mathbf{\Sigma}_Q = \tau\mathbf{\Pi}_Q$ we used to explain γ and μ, as it does not allow $\mathbf{\Sigma}_Q$ and $\boldsymbol{\beta}^*$ to be correlated.

4.3 Main Result

We are now ready to state and discuss our main result. For the proof as well as a more general result without Assumption A3 that covers arbitrary deterministic $(\mathbf{\Sigma}_P, \mathbf{\Sigma}_Q, \boldsymbol{\beta}^*)$, see Theorems 13 and 15 in Appendix D.

Theorem 3 (Monotonic risk relations) *Under Assumption A, the following hold with probability 1 in the limit as $d \to \infty$ for all $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathcal{D}_n, \ell, \lambda)$ solving* (1).

(a) *Regression. For $\psi(z^*, z) = (z^* - z)^2$, there exists a monotonic relation between $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}})$ and $\mathcal{R}_P(\widehat{\boldsymbol{\beta}})$ that depends only on $(P, Q, \boldsymbol{\beta}^*)$ if and only if Assumption B holds with $\gamma = \kappa$. If this relation exists, it is*

$$\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) = \gamma \mathcal{R}_P(\widehat{\boldsymbol{\beta}}) + \gamma r_P \sigma_\beta^2 (\mu - 1).$$

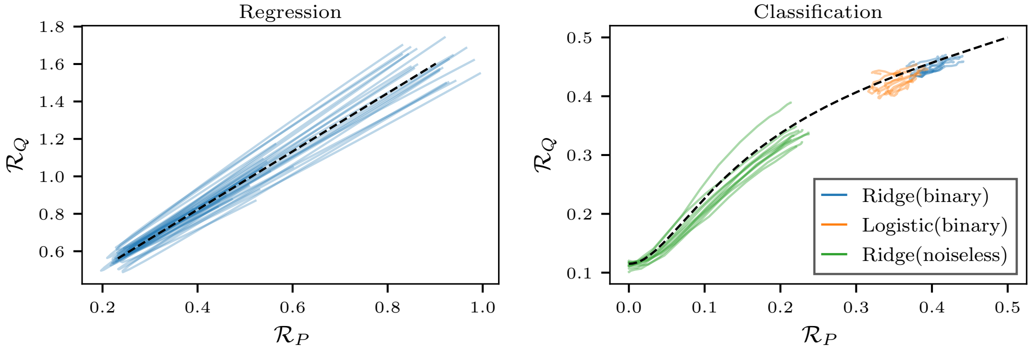


Figure 2: The risk relationships for data generated according to our distribution shift model match our theoretical results (dashed). Each colored curve corresponds to a sweep of the regularization strength of a single model on a single random trial. For both plots, we consider a subspace shift model with $\mathbf{\Sigma}_Q = \tau \mathbf{\Pi}_Q$ having $d_P/d = 0.9$, $d_Q/d = 0.8$, $d_{PQ}/d = 0.7$, and $\tau = 2$. We use $n = 1000$, $d = 800$, $\sigma_\beta^2 = 1$, and have $\gamma \approx 1.56$, and $\mu \approx 1.14$. **Left:** Mean squared error for ridge regression models (blue) trained on $y_i = \mathbf{x}_i^\top \beta^* + \sigma \xi_i$ for $\sigma^2 = 0.2$ and $\kappa = \gamma$. Although the tuning parameter overshoots the minimizer in the parameter sweep, it still always lies approximately on the line. **Right:** Misclassification error for ridge regression (blue) and logistic regression (orange) models with ridge penalty trained on corrupted binary labels generated as $\Pr(y_i = \operatorname{sign}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)) = 0.8$ with $\kappa = 5\gamma$. We also plot ridge regression trained on noiseless labels $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}^*$ (green) to illustrate that the result is independent of the labeling function, depending only on the feature distribution shift.

(b) *Classification. For* $\psi(z^*, z) = \mathbb{1}\{z^* z < 0\}$, *there exists a monotonic relation between* $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}})$ *and* $\mathcal{R}_P(\widehat{\boldsymbol{\beta}})$ *that depends only on* $(P, Q, \boldsymbol{\beta}^*)$ *if and only if Assumption B holds. If this relation exists, it is*

$$\sec^2(\pi \mathcal{R}_Q(\widehat{\boldsymbol{\beta}})) = \tfrac{\kappa\mu}{\gamma}\Big(\sec^2(\pi \mathcal{R}_P(\widehat{\boldsymbol{\beta}})) - 1\Big) + \mu,$$

where $\sec(t) = \frac{1}{\cos(t)}$. *Furthermore, if* $\mu = 1$, *then*

$$\log(\tan(\pi \mathcal{R}_Q(\widehat{\boldsymbol{\beta}}))) = \log(\tan(\pi \mathcal{R}_P(\widehat{\boldsymbol{\beta}}))) + \tfrac{1}{2}\log\tfrac{\kappa}{\gamma}.$$

Our result states that we have a monotonic relation between in- and out-of-distribution risks under our distribution shift model, for *all estimates* $\widehat{\boldsymbol{\beta}}(\mathcal{D}_n, \ell, \lambda)$ that solve a problem of the form (1), including, e.g., as we vary the training set size n, the regularization parameter λ, or even the labeling φ or loss function ℓ.

Figure 2 illustrates this behavior approximately in finite dimensions; there we plot the prediction of our theory along with realizations of data and estimates $\widehat{\boldsymbol{\beta}}(\mathcal{D}_n, \ell, \lambda)$. We see effects described by Theorem 3 in action: two models with the same risk on the distribution that generated the training data have the same risk on the new distribution, regardless of whether they were trained using regression or classification labels, of which particular loss function was used in training, of the training sample size, or of the level of label noise. As we can see in the figure, in finite dimensions individual models can have locally non-monotonic

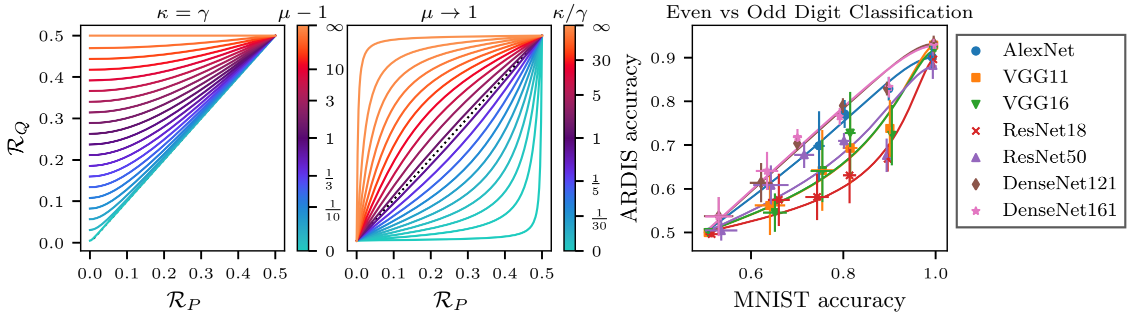


Figure 3: **Left/middle:** We plot the theoretical risk relation curves for misclassification error. General behavior of the risk relation is a combination of the two behaviors we demonstrate here. *Left:* $\mathcal{R}_Q$ converges to a nonzero limiting value as $\mathcal{R}_P \to 0$, which is determined by μ. *Middle:* While keeping the same limiting value of $\mathcal{R}_Q$ as $\mathcal{R}_P \to 0$, the shift is harder as κ/γ gets larger (red) and easier as κ/γ gets smaller (blue). **Right:** We train deep network models on classifying even vs. odd handwritten digits from the MNIST and ARDIS datasets, evaluating test performance during training as validation accuracy milestones are reached (dots with error bars over 8 trials). We also plot our theoretical risk relation with μ and κ/γ chosen to minimize squared error of the fit for each model.

relationships, and it is only as the system becomes asymptotically large and concentration of measure phenomena are realized that the monotonic relation emerges.

For both regression and classification, the risk relations are *linear*, with classification requiring the transformation $\mathcal{R} \mapsto \sec^2(\pi\mathcal{R})$ first before it becomes linear. This linearity is no coincidence; as we prove, whenever the risk depends linearly (after a fixed transformation) on some of the parameters of the covariances $\mathrm{Cov}(Z_P^*, Z_P)$ and $\mathrm{Cov}(Z_Q^*, Z_Q)$, as is the case for both squared error and misclassification error, the only monotonic relation that can exist is a linear one. We refer reader to Appendix D for proof details.

The risk relations are similar in that for both regression and classification, $\mu > 1$ indicates irreducible error due to a new subspace in Q that was unseen during training on P. However, the regression and classification risk relations also have a key difference: the squared error risk relation for regression only holds when $\gamma = \kappa$—i.e., only for task-independent shifts. This means that the subspace shift model with $\mathbf{\Sigma}_Q = \tau\mathbf{\Pi}_Q$ captures all aspects of the regression risk relation.

The classification risk relation, on the other hand, holds for task-dependent shifts with $\gamma \neq \kappa$. In particular, if we let $\mu \to 1$, then we find that the risk relation is remarkably similar to the empirical observation by Miller et al. (2021) that the risk relation is linear after applying an inverse Gaussian cumulative distribution function transformation $\Phi^{-1}(\cdot)$. Note that the $\log(\tan(\pi\cdot))$ transformation in Theorem 3 is strikingly similar to $\Phi^{-1}(\cdot)$; in fact, $\sup_{u\in\mathbb{R}} |\frac{1}{2}\Phi(u/\sqrt{2}) - \frac{1}{\pi}\tan^{-1}(e^u)| \leq 0.01$. This suggests that such "natural" distribution shifts formed by repeated dataset collection may have no subspace shift component ($\mu \to 1$), but rather only a task-dependent shift ($\gamma \neq \kappa$). We illustrate the behavior of the classification risk shift for different values of μ and κ/γ in Figure 3 (left).

For different feature spaces, our theory predicts different monotonic relations. This is also observed in practice: in Figure 3 (right), we show that except for the ResNet50 model, our theory predicts well the risk relation as a function of early stopping for deep network models trained on MNIST (LeCun et al., 2010), an easy handwritten digit classification task, and applied to ARDIS (Kusetogullari et al., 2020), a more difficult handwritten digits dataset. See Appendix B for a discussion of how this distribution shift fits the task-dependent shift model. The fits in Figure 3 show that different neural network models, which have their own respective implicit feature spaces, result in different monotonic risk relations. Because the shift is from an easy task to a hard one, we expect to see similar behavior to the case $\gamma < \kappa$, which matches the general trend of the fits, with some fits tending toward more or less task dependence based on model class. The tendency of models to dip in performance on ARDIS around 0.9 accuracy on MNIST is, we believe, a result of the change in the learned features of the networks during training, and is worst for ResNet50.

4.4 Settings without Monotonic Relations

Given the generality of the result in Theorem 3 across essentially any labeling function, training loss, and regularization strength, one might conjecture that the result holds for any risk and for any regularized ERM estimator. This is not the case, however, as the monotonic risk relations only arise due to the special structure of the risks and of ridge regularization.

As mentioned in the previous section, and as we elaborate on in the proof in Appendix D, monotonic risk relations arise when the metric ψ depends linearly on some one-dimensional function of the decision function covariances $\mathrm{Cov}(Z_P^*, Z_P)$ and $\mathrm{Cov}(Z_Q^*, Z_Q)$. The fact that squared error and misclassification error depend on different functions of the covariances is the first clue that the monotonicity of risk relations might not be universal. Indeed, the risk relations that arise are substantially distinct, as the misclassification relation captures task-dependent shifts while squared error does not.

As important counterexamples, popular convex losses used to train classification models such as the hinge loss and logistic loss do not exhibit monotonic risk relations. By Lemma 11, we know that the decision function covariances have only three degrees of freedom a, b, c (for general $\mathbf{\Sigma}_P$), but that the monotonic relation should hold regardless of how these are varied. In Figure 4, however, we show that as we vary even only a single parameter (here a), the hinge loss and logistic loss do *not* exhibit monotonic relations, while the misclassification error does. In general, monotonicity is further destroyed as we vary more degrees of freedom. This counterexample suggests that practitioners should be careful in their choice of validation metric: optimization of the in-distribution validation loss may not coincide with optimization of the out-of-distribution loss. Choosing a validation metric for which we expect monotonicity, such as misclassification error, is the better choice.

Another way that monotonicity can be broken is by changing the dependence of the decision function covariance on the underlying free parameters a, b, c. This occurs, for example, if we change the regularizer from the ridge penalty $\frac{1}{2}\|\cdot\|_2^2$ to some other regularizer such as the ℓ_1 norm $\|\cdot\|_1$. As we show in Appendix D.6, for separable regularizers, we still have monotonic relations, but now for only a restricted class of distributions shifts. Specifically, we only have monotonic relations in the task-independent setting ($\gamma = \kappa$), since in this case the covariances still admit similar linear decompositions. Otherwise, the nonlinearity due to the regularization penalty destroys monotonicity.

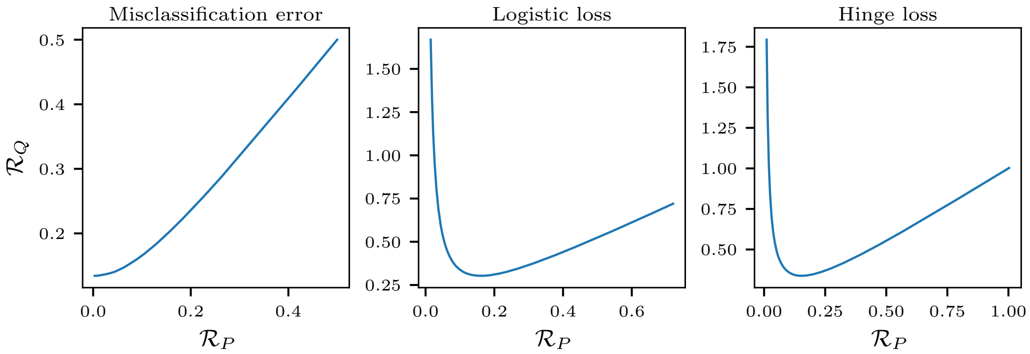


Figure 4: Using Monte Carlo simulation with 10^6 random draws of (Z^*, Z) from both P and Q under Assumption A, we compute the risk relationships for misclassification error (**left**) alongside the logistic loss $\psi(z^*, z) = \log(1+\exp(-\text{sign}(z^*)z))$ (**middle**) and the hinge loss $\psi(z^*, z) = \max\{1 - \text{sign}(z^*)z\}$ (**right**). Here we consider a subspace shift model with $d_P/d = 0.9$, $\sigma_\beta = 1$, $\gamma = 1$, $\kappa = 1$, $\mu = 1.2$. We fix degrees of freedom $b = c = 1$ and vary a. Unlike the misclassification error, these losses do not exhibit monotonic risk relationships as a function of a.

5. Linear Relations in Linear Inverse Problems

In this section, we switch to signal reconstruction problems, where linear relationships are also observed for signal reconstruction methods under distribution shifts (Darestani et al., 2021).

We consider a linear inverse problem setup where the measurement $\mathbf{y}$ is generated by a linear transform of the signal $\mathbf{x}$ plus some additive noise $\mathbf{z}$ independent of $\mathbf{x}$, i.e., $\mathbf{y} = \mathbf{A}\mathbf{x}+\mathbf{z}$, where $\mathbf{A} \in \mathbb{R}^{n\times d}$ with $n \leq d$, $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{z} \in \mathbb{R}^n$. We assume a similar signal subspace model as in Section 3: for data from distribution P, the signal is given by $\mathbf{x} = \mathbf{U}_P\mathbf{c}_P$, where $\mathbf{U}_P \in \mathbb{R}^{d\times d_P}$ has orthonormal columns and $\mathbf{c}_P \in \mathbb{R}^{d_P}$ is zero-mean and has identity covariance. The noise variable $\mathbf{z}$ is independent of $\mathbf{c}_P$ and has mean zero and covariance matrix $\sigma_P^2\mathbf{I}$. The data from distribution Q is generated in the same manner, but the signal is from a different subspace, i.e., $\mathbf{x} = \mathbf{U}_Q\mathbf{c}_Q$, where $\mathbf{U}_Q \in \mathbb{R}^{d\times d_Q}$ is orthonormal, $\mathbf{c}_Q \in \mathbb{R}^{d_Q}$ is zero-mean and has identity covariance, and the covariance matrix of the independent noise $\mathbf{z}$ is $\sigma_Q^2\mathbf{I}$. We assume that the number of measurements is larger than the subspace dimension, i.e., $d_P, d_Q \leq n$.

We consider the class of signal estimates given by

$$\widehat{\mathbf{x}}_\lambda(\mathbf{y}) = \mathbf{W}^*\mathbf{y}, \quad \mathbf{W}^* = \arg\min_{\mathbf{W}} \mathbb{E}_P\left[\|\mathbf{x} - \mathbf{W}\mathbf{y}\|_2^2\right] + \lambda\|\mathbf{W}\|_F^2.$$

Define the risk of an estimate $\widehat{\mathbf{x}}$ on distribution P with respect to the normalized squared error as $\mathcal{R}_P(\widehat{\mathbf{x}}) = \mathbb{E}_P\left[\left\|(\mathbf{x} - \widehat{\mathbf{x}})/\sqrt{d_P}\right\|_2^2\right]$ and likewise for distribution Q. We show that the relationship between $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ is captured by a similarity between subspaces $\mathbf{U}_P$ and $\mathbf{U}_Q$ that is determined by the principal angles between them. Let $\boldsymbol{\theta} \in \mathbb{R}^{\min\{d_P,d_Q\}}$ be the principal angles between subspaces spanned by the columns of $\mathbf{U}_P$ and $\mathbf{U}_Q$, and define $a = \|cos(\boldsymbol{\theta})\|_2^2/d_Q$.

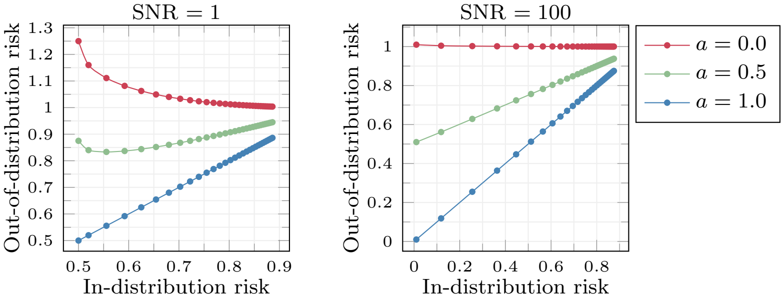


Figure 5: Non-linear relationship between risks $\mathcal{R}_P(\widehat{\mathbf{x}})$ and $\mathcal{R}_Q(\widehat{\mathbf{x}})$ in the low SNR regime (**left**) and approximate linear relationship in the high SNR regime (**right**) in signal denoising. Each curve plots the risks of estimate $\widehat{\mathbf{x}}_\lambda$ as the parameter λ varies. The signal-to-noise ratio is defined as $\text{SNR} = 1/\sigma_P^2$ and we set $\sigma_Q^2 = \sigma_P^2$. Shifts in the subspace is captured by $a = \|cos(\boldsymbol{\theta})\|_2^2/d_Q$.

Denoising. We start with denoising where the measurement matrix is the identity, i.e., $\mathbf{A} = \mathbf{I}$. It can be shown that $\widehat{\mathbf{x}}_\lambda(\mathbf{y}) = \alpha \mathbf{U}_P \mathbf{U}_P^\top \mathbf{y}$,where $\alpha = 1/(1 + \sigma_P^2 + \lambda)$. The following relationship between the risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ holds.

Theorem 4 *The risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ of the signal estimate $\widehat{\mathbf{x}}_\lambda$ obey*

$$\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) = a\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) + (1-a) + \alpha^2 \left(\frac{d_P}{d_Q}\sigma_Q^2 - a\sigma_P^2 \right),$$

where $\alpha = 1/(1 + \sigma_P^2 + \lambda)$.

In general, the relationship between the risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ is non-linear: it can be shown that $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) = (1-\alpha)^2 + \alpha^2\sigma_P^2$ (see the proof of Theorem 4), hence the term $\alpha^2 \left((d_P/d_Q)\sigma_Q^2 - a\sigma_P^2 \right)$ is not a linear function of the risk $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$. However, if the noise variances $\sigma_P^2, \sigma_Q^2 \ll 1$, then an approximate linear relation $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) \approx a\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) + (1-a)$ holds.

We illustrate Theorem 4 through a denoising simulation. In Figure 5, we plot the trajectory $(\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda), \mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda))$, as the parameter λ of the estimate $\widehat{\mathbf{x}}_\lambda$ varies. For high SNR the relationship between $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ is approximately linear, and for low SNR it is highly nonlinear.

Compressed sensing. We continue with compressed sensing, where the matrix $\mathbf{A}$ is a random matrix that down-samples the signal $\mathbf{x}$. Now the estimate $\widehat{\mathbf{x}}_\lambda(\mathbf{y})$ is only approximately $\alpha \mathbf{U}_P \mathbf{U}_P^\top \mathbf{y}$ due to the random measurement process. However, a similar relationship still holds between the risks.

Theorem 5 *Let $\mathbf{A} \in \mathbb{R}^{n \times d}$ be a random Gaussian matrix with independent entries drawn from the distribution $\mathcal{N}(0, 1/n)$. There exists a constant $c > 0$ such that, for any $0 < \epsilon < 1/d_P$, with probability at least $1 - 4(d_P(d_P + d_Q))\exp(-n\epsilon^2/8)$, it holds that*

$$\left| \mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) - a\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) - (1-a) - \alpha^2 \left(\frac{d_P}{d_Q}\sigma_Q^2 - a\sigma_P^2 \right) \right| \leq c\epsilon,$$

where $\alpha = 1/(1 + \sigma_P^2 + \lambda)$.

In the high SNR regime, if the number of measurements n is large enough, then with high probability there is an approximate linear relationship between the risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$.

6. Conclusion

In this paper, we studied the performance of estimators based on regularized empirical risk minimization trained on a distribution P, quantifying how they perform under distribution shifts on a distribution Q for regression, classification, and signal estimation problems. We identified conditions under which monotonic relations between the in-distribution risk $\mathcal{R}_P$ and out-of-distribution risk $\mathcal{R}_Q$ arise that hold for broad classes of regularized estimators, similarly to the linear risk relationships observed in practice.

Our findings in this work suggest that the linear and monotonic relations under distribution shifts observed in practice are emergent phenomena that arise from concentration of measure effects in large systems, which reduce the dependence of the risks down to only a single parameter. By identifying necessary and sufficient conditions for monotonic risk relations to exist, and characterizing the form of the monotonic relations, our work enables the principled discussion and investigation of such risk relations in future work.

Acknowledgments

This work was supported by the Institute of Advanced Studies at the Technical University of Munich, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 456465471, 464123524, and the authors also received funding by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. DL was also partially supported by the Office of Naval Research grant N00014-18-1-2047 and MURI grant N00014-20-1-2787 and the Army Research Office grant 2003514594.

RH would like to thank Ludwig Schmidt for helpful discussions on linear fits, and for feedback on the manuscript.

Appendix A. Details on the Experimental Results

Here, we provide further details on the numerical experiments in the main body.

A.1 Experimental Details for Object Detection

In this section, we describe the details of the object detection experiment from Section 2.

The models we evaluate are from `torchvision.models` and public github repositories:

- RetinaNet (Lin et al., 2017): RetinaNet ResNet-50 FPN
- Mask R-CNN (He et al., 2017): Mask R-CNN ResNet-50 FPN
- SSD (Liu et al., 2016): SSD300 VGG16, SSDlite320 MobileNetV3-Large
- Faster R-CNN (Ren et al., 2015): Faster R-CNN ResNet-50 FPN, Faster R-CNN MobileNetV3-Large FPN, Faster R-CNN MobileNetV3-Large 320 FPN

- Keypoint R-CNN (He et al., 2017): Keypoint R-CNN ResNet-50 FPN
- YOLOv5 (Redmon et al., 2016; Jocher et al., 2020): YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x

These model are trained on the COCO 2017 training set (Lin et al., 2014). We take the trained models and evaluate their performances on the COCO 2017 validation set and the VOC 2012 training/validation set (Everingham et al., 2010). Instead of using the standard metric for object detection—the mean average precision (mAP), which is the area under the precision-recall curve averaged over all classes—we consider the mean squared error in bounding box coordinates and only the `person` class. The predicted and the ground truth bounding box coordinates are normalized by the height and width of individual image. All models are evaluated using an NVIDIA A40 GPU.

To analyze the spectrum of the feature space of YOLOv5, we collect feature vectors through the following procedure. For each image in each evaluation set, we record the ground truth `person` objects that are correctly detected by *all* models listed above with an IOU threshold greater than or equal to 0.2. Then for each commonly detected ground truth object, we consider the prediction that has the largest IOU with the ground truth bounding box as the true positive. We then extract the feature vectors corresponding to these true positive predictions from the 24th layer of YOLOv5. This procedure is illustrated in Figure 6.

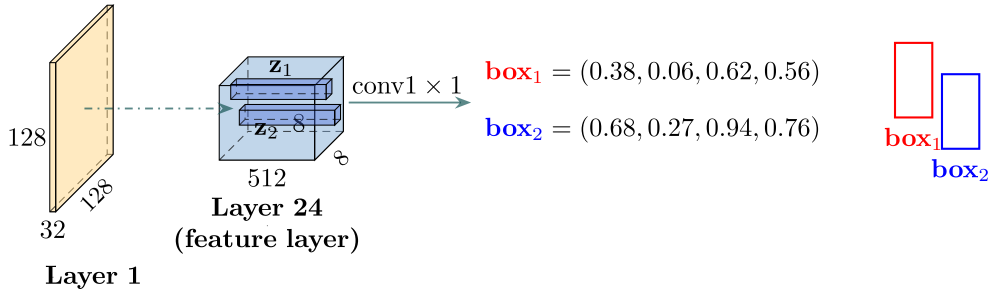


Figure 6: Visualization of feature extraction from YOLOv5: only feature vectors that correspond to true positive predictions are recorded for feature space analysis. The prediction of box_1, which is based on the feature vector $\mathbf{z}_1$, is true positive, while the prediction of box_2, which is based on the feature vector $\mathbf{z}_2$, is false positive. We record the feature vector $\mathbf{z}_1$ and discard the feature vector $\mathbf{z}_2$. Similarly, predictions in other grid positions in the 8×8 grid of the feature layer are not recorded if they do not correspond to a true positive prediction, since these feature vectors do not contain much information about the correct bounding box coordinates.

At the end, relevant feature vectors across the same evaluation set are stacked together and we obtain two sets of feature vectors $\mathcal{Z}_{\text{in}} = \{\mathbf{z}_j^{(i)} \in \mathbb{R}^{512} : i \in [N_{\text{in}}], j \in [K_{\text{in}}^{(i)}]\}$ and $\mathcal{Z}_{\text{out}} = \{\mathbf{z}_j^{(i)} \in \mathbb{R}^{512} : i \in [N_{\text{out}}], j \in [K_{\text{out}}^{(i)}]\}$ for the COCO 2017 and VOC 2012 evaluation dataset respectively, where $\mathbf{z}_j^{(i)}$ is the j^{th} true positive prediction on image i, N_{in} and N_{out}

are the numbers of images of the respective dataset and $K_{\text{in}}^{(i)}$ and $K_{\text{out}}^{(i)}$ are the number of true positive predictions on the i^{th} image respectively.

We make a few comments:

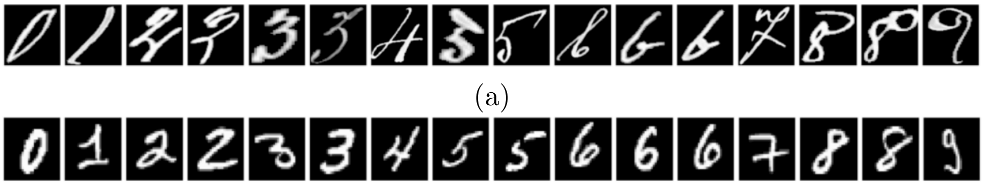
1 We only consider *common true positive* predictions: (1) for false positive and true negative predictions, there is no object to predict, hence the feature vectors contain no information for the regression task; (2) for false negative predictions, either the squared errors of the predicted coordinates are large since the IOU is lower than the 0.2 threshold, or they have lower confidence than another prediction which is true positive, so we simply exclude the corresponding feature vectors as they do not provide much useful information; (3) only common true positive predictions are considered so that all models make predictions on the same set of feature vectors.

2 YOLOv5 uses multiple layers (the 18th and 21st layers in addition to the 24th layer) as input to the bounding box prediction layer, but we find that most common true positive predictions are based on the 24th layer, probably due to the fact that this layer has a spacial dimension 8 × 8, where most ground truth objects size fit into, while the other layers have special dimension 16×16 and 32×32 matching small and tiny objects, which are relatively harder to predict.

A.2 Experimental Details for Digit Classification

In this section, we describe the details of the even vs odd handwritten digit classification experiment in Figure 3 (right).

We consider a binary classification task of classifying even versus odd digits on the MNIST (LeCun et al., 2010) dataset and ARDIS (Kusetogullari et al., 2020) dataset IV. The ARDIS dataset is a new image-based handwritten historical digit dataset extracted from Swedish church records, which induces a natural distribution shift from the widely used MNIST dataset. The ARDIS dataset IV has the same image size as the MNIST dataset with white digits in black background. The following figure shows examples of digits from both datasets.



(a)

(b)

Figure 7: Examples of digits: (a) ARDIS and (b) MNIST.

The models we evaluate are from `torchvision.models`:

- AlexNet (Krizhevsky et al., 2012)
- VGG (Simonyan and Zisserman, 2015): VGG11, VGG16
- ResNet (He et al., 2016): ResNet18, ResNet50
- DenseNet (Huang et al., 2017): DenseNet121, DenseNet161

Since the models we evaluate are originally designed for ImageNet (Deng et al., 2009) classification where the image sizes are larger, we resize the MNIST and ARDIS digits from 28×28 to 75×75. We train the model listed above on MNIST training set using the Adam optimizer with an initial learning rate 10^{-4} and a batch size 10 and a learning rate scheduler with a step size 10 epochs and a learning rate decay factor 0.1. The models at the top right corner of Figure 3(right) are trained for 20 epochs. Intermediate models are obtained by early stopping when validation accuracy first reaches $0.5, 0.6, 0.7, 0.8$ and 0.9. Each model is trained eight times with random initialization and with random shuffling of the training data, using different random seeds. All models are trained on an NVIDIA A40 GPU.

Appendix B. Intuition Regarding Feature Scaling

In both regression and classification, we find that the risk relations depend on the scaling of the features. We give intuition regarding where these can be seen in practice for two settings of real data that we consider in this paper.

Task-independent feature scaling. The scaling of the features may be uncorrelated with the ground truth coefficients $\boldsymbol{\beta}^*$, such as in the subspace shift case $\boldsymbol{\Sigma} = \rho \boldsymbol{\Pi}_Q$. An example in real data is that the principal components of the learned feature space of YOLOv5 model on VOC 2012 have uniformly larger magnitudes than those on COCO 2017, as can be seen from Figure 1.

Task-dependent feature scaling. The scaling of the features may be correlated with the ground truth coefficients $\boldsymbol{\beta}^*$. A motivating example in real data is the MNIST and ARDIS handwritten digit datasets. The universal ground truth labeling function for both of these datasets is the same (humans can classify digits from both datasets very well) and conceivably relies on a complex combination of features involving stroke and loop placement. Such features are for example the types of features found by nonlinear embedding techniques such as Isomap Tenenbaum et al. (2000). While both strokes and loops are present in both datasets, we observe that some of these occur with more frequency and intensity in one dataset versus the other. For example, italics and embellishments are much more common in ARDIS than in MNIST, as can be seen from Figure 7. We imagine that in feature space, this corresponds to a larger scaling of these features.

Appendix C. Proof of Theorem 1

In this section, we prove the main results on linear relations in linear regression in finite dimensions.

The proof of Theorem 1 is based on the following sufficient and necessary condition for a linear relationship between risks $\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)$.

Theorem 6 (Sufficient and necessary condition) *The risk $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)$ of estimator $\widehat{\boldsymbol{\beta}}_\lambda$ is an affine function of $\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)$, i.e., there exist a and b such that $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda) = a\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda) + b$, if and only if*

$$\boldsymbol{\beta}^{*\mathsf{T}} \boldsymbol{\Pi}_P \boldsymbol{\Sigma}_Q \boldsymbol{\Pi}_{P^\perp} \boldsymbol{\beta}^* = 0,$$

where $\mathbf{\Pi}_P = \mathbf{U}_P \mathbf{U}_P^\mathsf{T}$ *is a orthonormal projection onto the subspace spanned by the orthonormal matrix* $\mathbf{U}_P$*,* $\mathbf{\Pi}_{P^\perp} = \mathbf{I} - \mathbf{\Pi}_P$ *is a projection onto the orthogonal complement, and* $\mathbf{\Sigma}_Q = \mathbb{E}_Q \left[\mathbf{x}\mathbf{x}^\mathsf{T}\right]$*.*

If condition (a) holds, then $\mathbf{\Sigma}_Q$ and $\mathbf{\Pi}_P$ commute, and therefore $\mathbf{\Pi}_P \mathbf{\Sigma}_Q \mathbf{\Pi}_{P^\perp} = \mathbf{\Sigma}_Q \mathbf{\Pi}_P \mathbf{\Pi}_{P^\perp} = \mathbf{0}$. If condition ($b$) holds, then $\mathbf{\Pi}_{P^\perp} \boldsymbol{\beta}^{*\mathsf{T}} = \mathbf{0}$. In both cases, the term marked with $(*)$ in the proof of Theorem 6 becomes zero and the linear relationship $\mathcal{R}_Q(\boldsymbol{\beta}) = a\mathcal{R}_P(\boldsymbol{\beta}) + b$ holds with slope and intercept

$$\begin{aligned} a &= \frac{\boldsymbol{\beta}^{*\mathsf{T}} \mathbf{\Pi}_P \mathbf{\Sigma}_Q \mathbf{\Pi}_P \boldsymbol{\beta}^*}{\boldsymbol{\beta}^{*\mathsf{T}} \mathbf{\Pi}_P \boldsymbol{\beta}^*}, \\ b &= \boldsymbol{\beta}^{*\mathsf{T}} \left(\mathbf{\Sigma}_Q - a\mathbf{\Pi}_P\right) \boldsymbol{\beta}^* + \sigma_Q^2 - a\sigma_P^2. \end{aligned}$$

If condition (c) holds, then expectation of the term marked with $(*)$ is zero:

$$\mathbb{E}_{\boldsymbol{\beta}^*} \left[\boldsymbol{\beta}^{*\mathsf{T}} \mathbf{\Pi}_P \mathbf{\Sigma}_Q \mathbf{\Pi}_{P^\perp} \boldsymbol{\beta}^*\right] = \operatorname{tr}\left(\mathbf{\Pi}_P \mathbf{\Sigma}_Q \mathbf{\Pi}_{P^\perp} \mathbb{E}_{\boldsymbol{\beta}^*} \left[\boldsymbol{\beta}^* \boldsymbol{\beta}^{*\mathsf{T}}\right]\right) = \operatorname{tr}\left(\mathbf{\Sigma}_Q \mathbf{\Pi}_{P^\perp} \mathbf{\Pi}_P\right) = 0,$$

and $\mathbb{E}_{\boldsymbol{\beta}^*} \left[\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}_\lambda)\right] = a\mathbb{E}_{\boldsymbol{\beta}^*} \left[\mathcal{R}_P(\widehat{\boldsymbol{\beta}}_\lambda)\right] + b$ with slope and intercept

$$\begin{aligned} a &= \frac{\operatorname{tr}\left(\mathbf{\Pi}_P \mathbf{\Sigma}_Q \mathbf{\Pi}_P\right)}{\operatorname{tr}\left(\mathbf{\Pi}_P\right)}, \\ b &= \operatorname{tr}\left(\mathbf{\Sigma}_Q - a\mathbf{\Pi}_P\right) + \sigma_Q^2 - a\sigma_P^2. \end{aligned}$$

This concludes the proof of Theorem 1. It remains to prove Theorem 6.

Proof of Theorem 6. The idea is to relate the risks $\mathcal{R}_P(\boldsymbol{\beta})$ and $\mathcal{R}_Q(\boldsymbol{\beta})$ with the help of the parameter α. We start by expressing the risk of $\boldsymbol{\beta} = \alpha \mathbf{U}_P \mathbf{U}_P^\mathsf{T} \boldsymbol{\beta}^*$ on distribution P as a function of α

$$\begin{aligned} \mathcal{R}_P(\boldsymbol{\beta}) &= \mathbb{E}_P \left[(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta})^2\right] \\ &= (\boldsymbol{\beta}^* - \boldsymbol{\beta})^\mathsf{T} \mathbb{E}_P \left[\mathbf{x}\mathbf{x}^\mathsf{T}\right] (\boldsymbol{\beta}^* - \boldsymbol{\beta}) + \sigma_P^2 \\ &= \boldsymbol{\beta}^{*\mathsf{T}} (\mathbf{I} - \alpha \mathbf{U}_P \mathbf{U}_P^\mathsf{T}) \mathbf{U}_P \mathbb{E}_P \left[\mathbf{c}_P \mathbf{c}_P^\mathsf{T}\right] \mathbf{U}_P^\mathsf{T} (\mathbf{I} - \alpha \mathbf{U}_P \mathbf{U}_P^\mathsf{T}) \boldsymbol{\beta}^* + \sigma_P^2 \\ &= \boldsymbol{\beta}^{*\mathsf{T}} \mathbf{U}_P \mathbb{E}_P \left[\mathbf{c}_P \mathbf{c}_P^\mathsf{T}\right] \mathbf{U}_P^\mathsf{T} \boldsymbol{\beta}^* + (\alpha^2 - 2\alpha) \boldsymbol{\beta}^{*\mathsf{T}} \mathbf{U}_P \mathbb{E}_P \left[\mathbf{c}_P \mathbf{c}_P^\mathsf{T}\right] \mathbf{U}_P^\mathsf{T} \boldsymbol{\beta}^* + \sigma_P^2 \\ &= \boldsymbol{\beta}^{*\mathsf{T}} \mathbf{\Pi}_P \boldsymbol{\beta}^* + (\alpha^2 - 2\alpha) \boldsymbol{\beta}^{*\mathsf{T}} \mathbf{\Pi}_P \boldsymbol{\beta}^* + \sigma_P^2. \end{aligned}$$

Similarly, on distribution Q

$$\begin{aligned}
\mathcal{R}_Q(\boldsymbol{\beta}) &= \mathbb{E}_Q\left[(y - \mathbf{x}^\mathsf{T}\boldsymbol{\beta})^2\right] \\
&= (\boldsymbol{\beta}^* - \boldsymbol{\beta})^\mathsf{T}\mathbb{E}_Q\left[\mathbf{x}\mathbf{x}^\mathsf{T}\right]((\boldsymbol{\beta}^* - \boldsymbol{\beta})) + \sigma_Q^2 \\
&= \boldsymbol{\beta}^{*\mathsf{T}}(\mathbf{I} - \alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T})\boldsymbol{\Sigma}_Q(\mathbf{I} - \alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T})\boldsymbol{\beta}^* + \sigma_Q^2 \\
&= \boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}_Q\boldsymbol{\beta}^* + \alpha^2\boldsymbol{\beta}^{*\mathsf{T}}\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\Sigma}_Q\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\beta}^* \\
&\qquad - 2\alpha\boldsymbol{\beta}^{*\mathsf{T}}\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\Sigma}_Q(\mathbf{U}_P\mathbf{U}_P^\mathsf{T} + \mathbf{U}_{P\perp}\mathbf{U}_{P\perp}^\mathsf{T})\boldsymbol{\beta}^* + \sigma_Q^2 \\
&= \boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}_Q\boldsymbol{\beta}^* + (\alpha^2 - 2\alpha)\boldsymbol{\beta}^{*\mathsf{T}}\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\Sigma}_Q\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\beta}^* \\
&\qquad - 2\alpha\boldsymbol{\beta}^{*\mathsf{T}}\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\boldsymbol{\Sigma}_Q\mathbf{U}_{P\perp}\mathbf{U}_{P\perp}^\mathsf{T}\boldsymbol{\beta}^* + \sigma_Q^2 \\
&= \boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Sigma}_Q\boldsymbol{\beta}^* + (\alpha^2 - 2\alpha)\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Pi}_P^\mathsf{T}\boldsymbol{\Sigma}_Q\boldsymbol{\Pi}_P\boldsymbol{\beta}^* \\
&\qquad - 2\alpha\underbrace{\boldsymbol{\beta}^{*\mathsf{T}}\boldsymbol{\Pi}_P\boldsymbol{\Sigma}_Q\boldsymbol{\Pi}_{P\perp}\boldsymbol{\beta}^*}_{(*)} + \sigma_Q^2.
\end{aligned}$$

Since the risk $\mathcal{R}_P(\boldsymbol{\beta})$ depends linearly on $\alpha^2 - 2\alpha$, a linear relationship between $\mathcal{R}_Q(\boldsymbol{\beta})$ and $\mathcal{R}_P(\boldsymbol{\beta})$ is equivalent to $\mathcal{R}_Q(\boldsymbol{\beta})$ being also linearly dependent on $\alpha^2 - 2\alpha$. Hence, it is sufficient and necessary that the term marked with $(*)$ is zero.

Appendix D. Proof of Theorem 3

The proof of Theorem 3 involves the following steps:

(Step 1) *Asymptotics.* We invoke the result of Loureiro et al. (2021) (Theorem 9) to characterize the covariances of the decision functions of the estimator and ground truth in terms of three parameters (Lemma 11).

(Step 2) *Monotonicity for linear risks.* We prove a generic result for any risk that is parameterized as an affine function of some of its parameters, providing necessary and sufficient conditions for a monotonic relation (Lemma 12).

(Step 3) *Specific metrics.* We apply the generic result in Lemma 12 to squared error and misclassification error to obtain the most general results (Theorems 13 and 15).

(Step 4) *Simplifying assumptions.* To aid in interpretability, we apply Assumption A3 to simplify the necessary and sufficient conditions.

D.1 Step 1: Asymptotics

Assumption A4 states that the loss function is pseudo-Lipschitz continuous of order 2, which is defined as follows.

Definition 7 (Pseudo-Lipschitz continuity) *For a given $p \geq 1$, a function $\mathbf{f}\colon \mathbb{R}^r \to \mathbb{R}^s$ is called pseudo-Lipschitz of order p if there exists a constant $C > 0$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^r$,*

$$\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\| \leq C\|\mathbf{x}_1 - \mathbf{x}_2\|\left(1 + \|\mathbf{x}_1\|^{p-1} + \|\mathbf{x}_2\|^{p-1}\right).$$

We also need the definition of the proximal operator of a function.

Definition 8 (Proximal operator) *The proximal operator of a function $f\colon \mathbb{R}^r \to \mathbb{R}$ is the unique minimizer of the following objective:*

$$\mathrm{Prox}_f\,(\mathbf{z}) \triangleq \arg\min_{\mathbf{x}} f(\mathbf{x}) + \tfrac{1}{2}\|\mathbf{x}-\mathbf{z}\|_2^2.$$

We finally replace Assumption A with a slightly more general version, that implies Assumption A.

Assumption A* (General setup) *Assumption A holds with Assumption A3 replaced as follows.*

(A3*) *The ground truth coefficients $\boldsymbol{\beta}^*$ are deterministic, or they are random with sub-Gaussian one-dimensional marginals independent of $\mathcal{D}$, and the spectral distribution of $\boldsymbol{\Sigma}_P$ converges with bounded eigenvalues, such that $\frac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\boldsymbol{\beta}^*$ and $\frac{1}{d}\|\boldsymbol{\beta}^*\|_2^2$ converge to finite nonzero limits as $d \to \infty$.*

Armed with Assumption A*, we are now ready to re-state Theorem 5 of Loureiro et al. (2021) in our notation.

Theorem 9 *Under Assumption A*, there exist scalar coefficients $a \in \mathbb{R}$, $b, c, C_1, C_2, C_3 > 0$ such that for any pseudo-Lipschitz function $h\colon \mathbb{R}^d \to \mathbb{R}$ of order 2 and any $0 < \epsilon < C_1$, with probability at least $1 - \frac{C_2}{\epsilon^2}e^{-C_3 n\epsilon^4}$, the estimator $\widehat{\boldsymbol{\beta}}$ in* (1) *satisfies*

$$\left| h\left(\tfrac{1}{\sqrt{d}}\widehat{\boldsymbol{\beta}}\right) - h\left(\tfrac{1}{\sqrt{d}}\boldsymbol{\Sigma}_P^{-1/2}\mathrm{Prox}_{\frac{1}{b}\frac{1}{2}\left\|\boldsymbol{\Sigma}_P^{-1/2}\cdot\right\|_2^2}\left(\tfrac{a}{b}\boldsymbol{\Sigma}_P^{1/2}\boldsymbol{\beta}^* + \tfrac{\sqrt{c}}{b}\mathbf{g}\right)\right)\right| < \epsilon,$$

where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ is independent of $\boldsymbol{\beta}^*$.

Combining this theorem with

$$\mathrm{Prox}_{\frac{1}{2b}\left\|\boldsymbol{\Sigma}_P^{-1/2}\cdot\right\|_2^2}(\mathbf{z}) = \left(\mathbf{I}_d + \tfrac{1}{b}\boldsymbol{\Sigma}_P^{-1}\right)^{-1}\mathbf{z}$$

and using the Borel–Cantelli lemma, extending from a single function h to a sequence of functions that are uniformly pseudo-Lipschitz of order 2, we obtain the following corollary.

Corollary 10 *Under Assumption A*, there exist $a \in \mathbb{R}$, $b, c > 0$ such that for any pseudo-Lipschitz functions $h_d\colon \mathbb{R}^d \to \mathbb{R}$ of order 2 with uniform constant $C > 0$, the following holds almost surely for the estimator $\widehat{\boldsymbol{\beta}}$ in* (1)*:*

$$\lim_{d\to\infty} h_d\left(\tfrac{1}{\sqrt{d}}\widehat{\boldsymbol{\beta}}\right) = \lim_{d\to\infty} h_d\left(\tfrac{1}{\sqrt{d}}\boldsymbol{\Sigma}_P^{1/2}\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\left(a\boldsymbol{\Sigma}_P^{1/2}\boldsymbol{\beta}^* + \sqrt{c}\mathbf{g}\right)\right),$$

where $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_d)$ is independent of $\boldsymbol{\beta}^*$.

Finally, we obtain the form of the limiting covariances that we need for our proof.

Lemma 11 *Under Assumption A*, as $n, d \to \infty$, and assuming the limits below exist for any $a \in \mathbb{R}$, $b, c > 0$, there exist $a \in \mathbb{R}$, $b, c > 0$ such that the estimator in* (1) *has decision functions converging almost surely to $\widehat{Z}_P$ and $\widehat{Z}_Q$ that satisfy*

$$\mathbb{E}\left[Z_P^{*2}\right] = \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\boldsymbol{\beta}^*,$$

$$\mathbb{E}\left[Z_P^*\widehat{Z}_P\right] = \lim_{d\to\infty} \tfrac{a}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P^2\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^*,$$

$$\mathbb{E}\left[\widehat{Z}_P^2\right] = \lim_{d\to\infty} \tfrac{a^2}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P^3\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\boldsymbol{\beta}^* + \tfrac{c}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_P^2\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\right]$$

$$\mathbb{E}\left[Z_Q^{*2}\right] = \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_Q\boldsymbol{\beta}^*,$$

$$\mathbb{E}\left[Z_Q^*\widehat{Z}_Q\right] = \lim_{d\to\infty} \tfrac{a}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^*,$$

$$\mathbb{E}\left[\widehat{Z}_Q^2\right] = \lim_{d\to\infty} \tfrac{a^2}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^* + \tfrac{c}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\right].$$

Proof The variances $\mathbb{E}\left[Z_P^{*2}\right]$ and $\mathbb{E}\left[Z_Q^{*2}\right]$ are simply defined as stated. For $\mathbb{E}\left[Z_P^*\widehat{Z}_P\right]$, observe that the decision functions $\mathbf{x}^\top\boldsymbol{\beta}^*$ and $\mathbf{x}^\top\widehat{\boldsymbol{\beta}}$ have correlation

$$\mathop{\mathbb{E}}_{\mathbf{x}\sim P}\left[(\mathbf{x}^\top\boldsymbol{\beta}^*)(\mathbf{x}^\top\widehat{\boldsymbol{\beta}})\right] = \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\widehat{\boldsymbol{\beta}}.$$

The functions in the sequence $h_d(\mathbf{u}) = \frac{1}{\sqrt{d}}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\mathbf{u}$ are uniformly Lipschitz since $\frac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\boldsymbol{\beta}^*$ converges and $\boldsymbol{\Sigma}_P$ has uniformly bounded eigenvalues almost surely, so we can apply Corollary 10 to obtain the stated result. Similarly, for $\mathbb{E}\left[\widehat{Z}_P^2\right]$, the functions $h_d(\mathbf{u}) = \mathbf{u}^\top\boldsymbol{\Sigma}_P\mathbf{u}$ are pseudo-Lipschitz continuous of order 2 with uniform constant C. The calculation is analogous for $\mathbb{E}\left[Z_Q^*\widehat{Z}_Q\right]$ and $\mathbb{E}\left[\widehat{Z}_Q^2\right]$, applying the functions $h_d(\mathbf{u}) = \frac{1}{\sqrt{d}}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_Q\mathbf{u}$ and $h_d(\mathbf{u}) = \mathbf{u}^\top\boldsymbol{\Sigma}_Q\mathbf{u}$, respectively. ■

D.2 Step 2: Monotonicity for Linear Risks

Proving necessary and sufficient conditions for arbitrary risks is not a trivial task. However, if the risk has a *linear* structure in some of the free parameters (perhaps after some invertible transformation), we can exploit this linearity to show that any risk relation must be affine.

Lemma 12 *Consider the following functions defined on $\mathcal{A} \times \mathcal{B}$ for open sets $\mathcal{A} \subseteq \mathbb{R}^{k_A}$ and $\mathcal{B} \subseteq \mathbb{R}^{k_B}$:*

$$R_P(\mathbf{a}, \mathbf{b}) = h(\mathbf{w}(\mathbf{a})^\top\mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b})) \quad \textit{and} \quad R_Q(\mathbf{a}, \mathbf{b}) = h(\mathbf{w}(\mathbf{a})^\top\mathbf{v}_Q(\mathbf{b}) + v_Q^0(\mathbf{b})),$$

where

- *$h\colon \mathbb{R} \to \mathbb{R}$ is a monotonically increasing or decreasing function,*
- *$(\mathbf{w}(\mathbf{a}), 1) \in \mathbb{R}^{k_W+1}$ is a vector of linearly independent scalar functions of $\mathbf{a}$ over $\mathcal{A}$,*

- $\mathbf{v}_P$, $\mathbf{v}_Q$, v_P^0, *and* v_Q^0 *are differentiable functions of* $\mathbf{b}$, *and* $\mathbf{v}_P(\mathbf{b}) \neq \mathbf{0}$ *for all* $\mathbf{b} \in \mathcal{B}$.

The following statements are equivalent:

(i) There exists a monotonically increasing function $u\colon \mathbb{R} \to \mathbb{R}$ *such that* $R_Q(\mathbf{a}, \mathbf{b}) = u(R_P(\mathbf{a}, \mathbf{b}))$ *for all* $\mathbf{a}, \mathbf{b} \in \mathcal{A} \times \mathcal{B}$.

(ii) There exists $\rho > 0$, $u_0 \in \mathbb{R}$ *such that for all* $\mathbf{b} \in \mathcal{B}$, $\mathbf{v}_Q(\mathbf{b}) = \rho \mathbf{v}_P(\mathbf{b})$ *and* $v_Q^0(\mathbf{b}) = \rho v_P^0(\mathbf{b}) + u_0$.

Furthermore, if u *exists, it has the form* $u(t) = h\left(\rho h^{-1}(t) + u_0\right)$.

Proof We first show that condition *(ii)* implies *(i)*. Denote $t(\mathbf{a}, \mathbf{b}) = \mathbf{w}(\mathbf{a})^\top \mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b})$ and note that condition *(ii)* implies that $R_P(\mathbf{a}, \mathbf{b}) = h(t(\mathbf{a}, \mathbf{b}))$ and $R_Q(\mathbf{a}, \mathbf{b}) = h(\rho t(\mathbf{a}, \mathbf{b}) + u_0)$. Next, note that the function $\tilde{u}(t) = \rho t + u_0$, $\rho > 0$ is monotonically increasing, and so is $u = h \circ \tilde{u} \circ h^{-1}$, since a composition of increasing and decreasing functions is increasing if the number of decreasing functions is even, and h and h^{-1} are either both increasing or both decreasing. Thus, condition *(ii)* implies *(i)*.

It remains to show that condition *(i)* implies *(ii)*. For this, we identify necessary conditions for *(i)* to hold. First note that by a similar argument to the *(ii)* $\implies$ *(i)* case , *(i)* holds if and only if there is a monotonic $\tilde{u}$ such that

$$\mathbf{w}(\mathbf{a})^\top \mathbf{v}_Q(\mathbf{b}) + v_Q^0(\mathbf{b}) = \tilde{u}(\mathbf{w}(\mathbf{a})^\top \mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b})). \tag{2}$$

In the following, we show that for this equation to hold, the function $\tilde{u}$ must have the form $\tilde{u}(t) = \rho t + u_0$ for $\rho > 0$.

We begin by taking the gradient of both sides of equation (2) with respect to $\mathbf{w}(\mathbf{a})$, giving the condition

$$\mathbf{v}_Q(\mathbf{b}) = \tilde{u}'(\mathbf{w}(\mathbf{a})^\top \mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b}))\mathbf{v}_P(\mathbf{b}). \tag{3}$$

Since the above equation must hold for all $\mathbf{a}, \mathbf{b} \in \mathcal{A} \times \mathcal{B}$, the derivative $\tilde{u}'\colon \mathbb{R} \to \mathbb{R}$ must be a function of $\mathbf{b}$ only—let us write this as $\rho(\mathbf{b}) \triangleq \tilde{u}'(\mathbf{w}(\mathbf{a})^\top \mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b}))$. We can additionally take the gradients of equation (2) with respect to $\mathbf{b}$:

$$\nabla_{\mathbf{b}} \mathbf{v}_Q(\mathbf{b}) \mathbf{w}(\mathbf{a}) + \nabla_{\mathbf{b}} v_Q^0(\mathbf{b}) = \tilde{u}'(\mathbf{w}(\mathbf{a})^\top \mathbf{v}_P(\mathbf{b}) + v_P^0(\mathbf{b})) \left(\nabla_{\mathbf{b}} \mathbf{v}_P(\mathbf{b}) \mathbf{w}(\mathbf{a}) + \nabla_{\mathbf{b}} v_P^0(\mathbf{b})\right).$$

We can rewrite this equation as

$$\left[\nabla_{\mathbf{b}}(\mathbf{v}_Q(\mathbf{b}), v_Q^0(\mathbf{b})) - \rho(\mathbf{b}) \nabla_{\mathbf{b}}(\mathbf{v}_P(\mathbf{b}), v_P^0(\mathbf{b}))\right](\mathbf{w}(\mathbf{a}), 1) = \mathbf{0}.$$

In this form, we can see that because $(w(\mathbf{a}), 1)$ is a vector of linearly independent functions over $\mathbf{a} \in \mathcal{A}$, the only solutions to this equation are the trivial solutions which satisfy

$$\nabla_{\mathbf{b}}(\mathbf{v}_Q(\mathbf{b}), v_Q^0(\mathbf{b})) = \rho(\mathbf{b}) \nabla_{\mathbf{b}}(\mathbf{v}_P(\mathbf{b}), v_P^0(\mathbf{b})). \tag{4}$$

Returning to equation (3), we can now take its gradient with respect to $\mathbf{b}$, yielding

$$\nabla_{\mathbf{b}} \mathbf{v}_Q(\mathbf{b}) = \left(\nabla_{\mathbf{b}} \rho(\mathbf{b})\right) \mathbf{v}_P(\mathbf{b})^\top + \rho(\mathbf{b}) \nabla_{\mathbf{b}} \mathbf{v}_P(\mathbf{b}),$$

which, combined with equation (4) implies that $(\nabla_{\mathbf{b}}\rho(\mathbf{b}))\,\mathbf{v}_P(\mathbf{b})^\top = \mathbf{0}$, implying that $\nabla_{\mathbf{b}}\rho(\mathbf{b}) = \mathbf{0}$ since $\mathbf{v}_P(\mathbf{b}) \neq \mathbf{0}$ by assumption. Thus, $\rho(\mathbf{b})$ is constant as a function of $\mathbf{b}$, implying that $\tilde{u}$ is an affine function; let us therefore write $\tilde{u}(t) = \rho t + u_0$. Then we can rewrite equation (2) as

$$\left[(\mathbf{v}_Q(\mathbf{b}), v_Q^0(\mathbf{b})) - \rho(\mathbf{v}_P(\mathbf{b}), v_P^0(\mathbf{b})) - (\mathbf{0}, u_0)\right]^\top (\mathbf{w}(\mathbf{a}), 1) = 0.$$

By linear independence again, this equation can have only the trivial solution, implying that $v_Q^0(\mathbf{b}) = \rho v_P^0(\mathbf{b}) + u_0$. Lastly, this mapping is monotonically increasing only if $\rho > 0$. ■

D.3 Step 3: Squared Error

We start with the simpler case of squared error. We first introduce notation to simplify expressions. Let

$$\begin{aligned}
\mathbb{E}\left[Z_P^{*2}\right] &= \Omega_P, \quad \mathbb{E}\left[Z_P^* \widehat{Z}_P\right] = a\Gamma_P(b), \quad \mathbb{E}\left[\widehat{Z}_P^2\right] = a^2\Lambda_P(b) + c\Theta_P(b), \\
\mathbb{E}\left[Z_Q^{*2}\right] &= \Omega_Q, \quad \mathbb{E}\left[Z_Q^* \widehat{Z}_Q\right] = a\Gamma_Q(b), \quad \mathbb{E}\left[\widehat{Z}_Q^2\right] = a^2\Lambda_Q(b) + c\Theta_Q(b),
\end{aligned}$$

where

$$\begin{gathered}
\Omega_P \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\boldsymbol{\beta}^*, \quad \Gamma_P(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P^2\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^*, \\
\Lambda_P(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P^3\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\boldsymbol{\beta}^*, \quad \Theta_P(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_P^2\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\right], \\
\Omega_Q \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_Q\boldsymbol{\beta}^*, \quad \Gamma_Q(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^*, \\
\Lambda_Q(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-1}\boldsymbol{\beta}^*, \\
\Theta_Q(b) \triangleq \lim_{d\to\infty} \tfrac{1}{d}\operatorname{tr}\left[\boldsymbol{\Sigma}_Q\boldsymbol{\Sigma}_P\left(\boldsymbol{\Sigma}_P + b\mathbf{I}_d\right)^{-2}\right].
\end{gathered} \tag{5}$$

We now prove the squared error case in the following theorem.

Theorem 13 *Under Assumption A*, with probability 1, in the limit as $d \to \infty$ for $\hat{f}(\mathbf{x}) = \phi(\mathbf{x}, \widehat{\boldsymbol{\beta}}(\mathcal{D}, \ell, \lambda))$ solving (1), for $\psi(z^*, \hat{z}) = (z^* - \hat{z})^2$, there exists a monotonic relation between $\mathcal{R}_Q(\hat{f})$ and $\mathcal{R}_P(\hat{f})$ that depends only on $(P, Q, \boldsymbol{\beta}^*)$ if and only if there exists $\rho > 0$ such that for all $b > 0$,*

$$\Gamma_Q(b) = \rho\Gamma_P(b), \quad \Lambda_Q(b) = \rho\Lambda_P(b), \quad \Theta_Q(b) = \rho\Theta_P(b).$$

If this relation exists, it is

$$\mathcal{R}_Q(\hat{f}) = \rho(\mathcal{R}_P(\hat{f}) - \Omega_P) + \Omega_Q.$$

Proof We begin by observing that

$$\mathcal{R}_P(\hat{f}) = \mathbb{E}\left[(Z_P^* - \widehat{Z}_P)^2\right] = \mathbb{E}\left[Z_P^{*2}\right] - 2\,\mathbb{E}\left[Z_P^*\widehat{Z}_P\right] + \mathbb{E}\left[\widehat{Z}_P^2\right],$$

which means that we can apply Lemma 12 with $h(t) = t$, $\mathbf{w}(a, c) = (-2a, a^2, c)$, and

$$\begin{aligned} \mathbf{v}_P(b) &= (\Gamma_P(b), \Lambda_P(b), \Theta_P(b)), \quad v_P^0(b) = \Omega_P, \\ \mathbf{v}_Q(b) &= (\Gamma_Q(b), \Lambda_Q(b), \Theta_Q(b)), \quad v_Q^0(b) = \Omega_Q. \end{aligned}$$

Therefore, the condition that $\mathbf{v}_Q(b) = \rho \mathbf{v}_P(b)$ is equivalent to the stated condition. The condition that $v_Q^0(b) = \rho v_P^0(b) + u_0$ is trivially satisfied by $u_0 = v_Q^0(b) - \rho v_P^0(b)$ since v_P^0 and v_Q^0 are constant functions of b. ■

D.4 Step 3: Misclassification Error

We now move to the slightly more difficult case of misclassification error. We first need a closed-form expression for the risk, which we obtain from the following lemma.

Lemma 14 *For two zero-mean jointly Gaussian random variables X and Y,*

$$\Pr(XY < 0) = \frac{1}{\pi} \arccos\left(\frac{\mathbb{E}\left[XY\right]}{\sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]}}\right).$$

Proof First define $\widetilde{X} = X/\sqrt{\mathbb{E}\left[X^2\right]}$ and $\widetilde{Y} = Y/\sqrt{\mathbb{E}\left[Y^2\right]}$. We can decompose $\widetilde{Y}$ as:

$$\widetilde{Y} = \mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]\widetilde{X} + \sqrt{1 - \mathbb{E}\left[\widehat{X}\widetilde{Y}\right]^2} U_Y,$$

where U_Y is a standard normal random variable. Observe that for any scalar $a > 0$, $\left\{\widetilde{X}\widetilde{Y} < 0\right\} = \left\{a\widetilde{X}\widetilde{Y} < 0\right\}$, so we can jointly scale $\widetilde{X}$ and U_Y without affecting the event, even if this scalar is random. Because $\widetilde{X}$ and U_Y are independent standard normal variables, this means we can choose a random variable $\Theta \sim \mathrm{Uniform}[0, 2\pi)$ such that

$$(\cos\Theta, \sin\Theta) = \left(\frac{\widetilde{X}}{\sqrt{\widetilde{X}^2 + U_Y^2}}, \frac{U_Y}{\sqrt{\widetilde{X}^2 + U_Y^2}}\right).$$

Now

$$\Pr(XY < 0) = \Pr\left(\widetilde{X}\widetilde{Y} < 0\right) = \Pr\left(\cos\Theta\left(\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]\cos\Theta + \sqrt{1 - \mathbb{E}\left[\widehat{X}\widetilde{Y}\right]^2}\sin\Theta\right) < 0\right).$$

This inequality is satisfied for

$$\Theta \in [0, 2\pi) \cap \bigcup_{n=-\infty}^{\infty}\left(\frac{(2n+1)\pi}{2}, \frac{(2n+1)\pi}{2} + \arccos\left(\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]\right)\right).$$

The size of each of the intervals in the union is $\arccos\left(\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]\right)$, and twice the length of one such interval is included in $[0, 2\pi)$. Plugging in the definitions of $\widetilde{X}$ and $\widetilde{Y}$ therefore proves the claim. ■

We are now ready to state and prove the classification error case.

Theorem 15 *Under Assumption A*, with probability 1, in the limit as $d \to \infty$ for $\hat{f}(\mathbf{x}) = \phi(\mathbf{x}, \widehat{\boldsymbol{\beta}}(\mathcal{D}, \ell, \lambda))$ solving (1), for $\psi(z^*, \hat{z}) = \mathbb{1}\{z^* \hat{z} < 0\}$, there exists a monotonic relation between $\mathcal{R}_Q(\hat{f})$ and $\mathcal{R}_P(\hat{f})$ that depends only on $(P, Q, \boldsymbol{\beta}^*)$ if and only there exist $\rho > 0$ and $u_0 \in \mathbb{R}$ such that for all $b > 0$*

$$\frac{\Omega_Q \Theta_Q(b)}{\Gamma_Q(b)^2} = \frac{\rho \Omega_P \Theta_P(b)}{\Gamma_P(b)^2}, \qquad \frac{\Omega_Q \Lambda_Q(b)}{\Gamma_Q(b)^2} = \frac{\rho \Omega_P \Lambda_P(b)}{\Gamma_P(b)^2} + u_0.$$

If this relation exists, it is

$$\sec^2(\pi \mathcal{R}_Q(\hat{f})) = \rho \sec^2(\pi \mathcal{R}_P(\hat{f})) + u_0,$$

where $\sec(t) = \frac{1}{\cos(t)}$.

Proof Let h have inverse $h^{-1}(t) = \sec^2(\pi t)$. Then applying Lemma 14 and the definitions in (5), the risk has the form

$$h^{-1}(\mathcal{R}_P(\hat{f})) = \frac{\mathbb{E}\left[Z_P^{*2}\right] \mathbb{E}\left[\widehat{Z}_P^2\right]}{\mathbb{E}\left[Z_P^* \widehat{Z}_P\right]^2} = \Omega_P \frac{a^2 \Lambda_P(b) + c\Theta_P(b)}{(a\Gamma_P(b))^2},$$

which means that we can apply Lemma 12 with $w(a, c) = \frac{c}{a^2}$ and

$$v_P(b) = \frac{\Omega_P \Theta_P(b)}{\Gamma_P(b)^2}, \quad v_Q(b) = \frac{\Omega_Q \Theta_Q(b)}{\Gamma_Q(b)^2}, \quad v_P^0(b) = \frac{\Omega_P \Lambda_P(b)}{\Gamma_P(b)^2}, \quad v_Q^0(b) = \frac{\Omega_Q \Lambda_Q(b)}{\Gamma_Q(b)^2}.$$

The condition from Lemma 12 is equivalent to the stated condition. ■

D.5 Step 4: Simplifying Assumptions

The necessary and sufficient conditions in Theorems 13 and 15 are rather difficult to interpret, and they do not simplify cleanly without additional assumptions. The strongest assumption we make is that $\boldsymbol{\Sigma}_P = \boldsymbol{\Pi}_P$ is a projection operator. The advantage of this is that it only has eigenvalues 0 and 1, which means that any term involving $(\boldsymbol{\Sigma}_P + b\mathbf{I}_d)^{-1}$ can have $\frac{1}{1+b}$ factored out, allowing all of the terms to simplify greatly. Because $\boldsymbol{\beta}^*$ has i.i.d. sub-Gaussian elements, we can without loss of generality assume it to be Gaussian having the same second moment and thus rotationally invariant. Combining these with the simultaneous diagonizability of $\boldsymbol{\Sigma}_Q$ and $\boldsymbol{\Sigma}_P$ gives us the following simplifications:

$$\Omega_P = r_P \sigma_\beta^2, \quad \Gamma_P(b) = \frac{r_P \sigma_\beta^2}{1+b}, \quad \Lambda_P(b) = \frac{r_P \sigma_\beta^2}{(1+b)^2}, \quad \Theta_P(b) = \frac{r_P}{(1+b)^2},$$
$$\Gamma_Q(b) = \lim_{d \to \infty} \frac{\boldsymbol{\beta}_P^{*\top} \boldsymbol{\Sigma}_Q \boldsymbol{\beta}_P^*}{d(1+b)}, \quad \Lambda_Q(b) = \lim_{d \to \infty} \frac{\boldsymbol{\beta}_P^{*\top} \boldsymbol{\Sigma}_Q \boldsymbol{\beta}_P^*}{d(1+b)^2}, \quad \Theta_Q(b) = \lim_{d \to \infty} \frac{\operatorname{tr}\left[\boldsymbol{\Sigma}_Q \boldsymbol{\Pi}_P\right]}{d(1+b)^2}.$$

For γ, κ, and μ from Assumption B, we therefore have the following relations:

$$\Gamma_Q(b) = \gamma \Gamma_P(b), \quad \Lambda_Q(b) = \gamma \Lambda_P(b), \quad \Theta_Q(b) = \kappa \Theta_P(b), \quad \Omega_Q = \gamma \mu \Omega_P,$$
$$\frac{\Omega_Q \Theta_Q(b)}{\Gamma_Q(b)^2} = \frac{\mu \kappa \Omega_P \Theta_P(b)}{\gamma \Gamma_P(b)^2}, \quad \frac{\Omega_Q \Lambda_Q(b)}{\Gamma_Q(b)^2} = \frac{\mu \Omega_P \Lambda_P(b)}{\Gamma_P(b)^2} = \mu = \frac{\mu \kappa}{\gamma} + \mu\left(1 - \frac{\kappa}{\gamma}\right).$$

For regression, this means that in Theorem 13, $\rho = \gamma = \kappa$, and $\Omega_Q - \rho\Omega_P = \gamma r_P \sigma_\beta^2(\mu - 1)$. For classification, this means that in Theorem 15, $\rho = \frac{\mu\kappa}{\gamma}$ and $u_0 = \mu(1 - \frac{\kappa}{\gamma})$. These values give the stated claims in Theorem 3, and when specializing to $\mu = 1$ for classification, the relation follows by the fact that $\tan^2(\theta) = \sec^2(\theta) - 1$.

D.6 General Regularization Penalties

The above approach can be used to analyze general separable regularization penalties as well via linearization if $\mathbf{\Sigma}_P$ is axis-aligned (that is, diagonal). Under Assumption A, the equations in Lemma 11 simplify to the forms shown in Step 4 of the proof of Theorem 3. Upon closer inspection, we observe that instead of three free variables a, b, c, we now only have two degrees of freedom via $\frac{a}{1+b}$ and $\frac{c}{1+b}$. Meanwhile, let

$$\widehat{\boldsymbol{\beta}}(\mathcal{D}, \ell, \lambda) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda \sum_{j=1}^{d} r([\boldsymbol{\beta}]_j)^2$$

for a convex regularization penalty $r\colon \mathbb{R} \to \mathbb{R}$. Corollary 10 can be extended to general regularization penalties (see Loureiro et al., 2021), and our resulting estimator has the following form for each $j \in [d]$ such that $[\mathbf{\Sigma}_P]_{jj} = 1$:

$$[\widehat{\boldsymbol{\beta}}]_j \simeq \operatorname{Prox}_{r(\cdot)/b}\left(\tfrac{a}{b}[\boldsymbol{\beta}^*]_j + \tfrac{\sqrt{c}}{b}[\mathbf{g}]_j\right)$$

for some three parameters $a \in \mathbb{R}$, $b, c > 0$. Assuming $r(u)$ is an increasing function of $|u|$, this implies that the remaining coefficients $\widehat{\boldsymbol{\beta}}$ will be 0.

As in the proof of Lemma 11, we only need to determine the following inner products:

$$\tfrac{1}{d}\boldsymbol{\beta}^{*\top}\mathbf{\Pi}_P\widehat{\boldsymbol{\beta}}, \quad \tfrac{1}{d}\widehat{\boldsymbol{\beta}}^\top\mathbf{\Pi}_P\widehat{\boldsymbol{\beta}}, \quad \tfrac{1}{d}\boldsymbol{\beta}^{*\top}\mathbf{\Sigma}_Q\widehat{\boldsymbol{\beta}}, \quad \tfrac{1}{d}\widehat{\boldsymbol{\beta}}^\top\mathbf{\Sigma}_Q\widehat{\boldsymbol{\beta}}.$$

For any $a \in \mathbb{R}$, $b, c > 0$, we can linearize $\widehat{\boldsymbol{\beta}}$ in the form $a'\boldsymbol{\beta}^* + \sqrt{c'}\mathbf{g}$ with respect to $\boldsymbol{\beta}^*$ and $\mathbf{\Pi}_P$ in the sense that we can find $a' \in \mathbb{R}, c' > 0$ such that

$$\tfrac{1}{d}\boldsymbol{\beta}^{*\top}\mathbf{\Pi}_P\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} a' r_P \sigma_\beta^2, \quad \tfrac{1}{d}\widehat{\boldsymbol{\beta}}^\top\mathbf{\Pi}_P\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} a'^2 r_P \sigma_\beta^2 + c' r_P,$$

which is the same as we have in the ridge regularization case. Therefore, Theorem 3 will also apply for an arbitrary separable regularizer if and only if

$$\tfrac{1}{d}\boldsymbol{\beta}^{*\top}\mathbf{\Sigma}_Q\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} a'\tfrac{1}{d}\boldsymbol{\beta}_P^{*\top}\mathbf{\Sigma}_Q\boldsymbol{\beta}_P^*,$$
$$\tfrac{1}{d}\widehat{\boldsymbol{\beta}}^\top\mathbf{\Sigma}_Q\widehat{\boldsymbol{\beta}} \xrightarrow{\text{a.s.}} a'^2\boldsymbol{\beta}_P^{*\top}\mathbf{\Sigma}_Q\boldsymbol{\beta}_P^* + c'\tfrac{1}{d}\operatorname{tr}[\mathbf{\Sigma}_Q\mathbf{\Pi}_P].$$

Due to the nonlinearity of the proximal operator, we would not expect these to hold in general, as our linearization only holds for $\widehat{\boldsymbol{\beta}}$ measured with respect to $\mathbf{\Pi}_P$. However, for example, if $\kappa = \gamma$, then the linearization holds and we can apply Theorem 3.

Appendix E. Proof of Theorem 4 and Theorem 5

In this section, we prove the main results on linear relations for linear inverse problems.

E.1 Proof of Theorem 4

The proof is similar to that of Theorem 6. We first express the risk of the signal estimate $\widehat{\mathbf{x}}_\lambda$ on distribution P as a function of $\alpha = 1/(1+\sigma_P^2+\lambda)$. Note that the solution $\mathbf{W}^*$ to $\min_{\mathbf{W}} \mathbb{E}_P \left[\|\mathbf{x}-\mathbf{W}\mathbf{y}\|_2^2\right] + \lambda\|\mathbf{W}\|_F^2$ can be computed as

$$\begin{aligned}
\mathbf{W}^* &= \mathbb{E}_P\left[\mathbf{x}\mathbf{y}^\mathsf{T}\right]\left(\mathbb{E}_P\left[\mathbf{y}\mathbf{y}^\mathsf{T}\right]+\lambda\mathbf{I}\right)^{-1} \\
&\overset{(a)}{=} \mathbb{E}_P\left[\mathbf{x}\mathbf{x}^\mathsf{T}\right]\left(\mathbb{E}_P\left[\mathbf{x}\mathbf{x}^\mathsf{T}+\mathbf{z}\mathbf{z}^\mathsf{T}\right]+\lambda\mathbf{I}\right)^{-1} \\
&\overset{(b)}{=} \mathbf{U}_P\mathbf{U}_P^\mathsf{T}\left(\mathbf{U}_P\mathbf{U}_P^\mathsf{T}+(\sigma_P^2+\lambda)\mathbf{I}\right)^{-1} \\
&= \frac{1}{1+\sigma_P^2+\lambda}\mathbf{U}_P\mathbf{U}_P^\mathsf{T},
\end{aligned}$$

where (a) follows from that $\mathbf{z}$ is independent of $\mathbf{x}$ and that $\mathbb{E}[\mathbf{z}] = 0$, and (b) follows from the assumptions that $\mathbb{E}_P\left[\mathbf{c}_P\mathbf{c}_P^\mathsf{T}\right] = \mathbf{I}$ and that $\mathbb{E}_P\left[\mathbf{z}\mathbf{z}^\mathsf{T}\right] = \sigma_P^2\mathbf{I}$. Hence, $\widehat{\mathbf{x}}_\lambda(\mathbf{y}) = \alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{y}$. It holds that

$$\begin{aligned}
\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) &= \mathbb{E}_P\left[\left\|(\mathbf{U}_P\mathbf{c}_P-\widehat{\mathbf{x}}_\lambda)/\sqrt{d_P}\right\|_2^2\right] \\
&= \mathbb{E}_P\left[\left\|\left(\mathbf{I}-\alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\right)\mathbf{U}_P\mathbf{c}_P/\sqrt{d_P}\right\|_2^2\right] + \mathbb{E}_P\left[\left\|\alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{z}/\sqrt{d_P}\right\|_2^2\right] \\
&= \mathbb{E}_P\left[\|(1-\alpha)\,\mathbf{U}_P\mathbf{c}_P\|_2^2/d_P\right] + \operatorname{tr}\left(\alpha^2\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbb{E}_P\left[\mathbf{z}\mathbf{z}^\mathsf{T}\right]/d_P\right) \\
&= \operatorname{tr}\left[(1-\alpha)^2\,\mathbf{U}_P^\mathsf{T}\mathbf{U}_P\,\mathbb{E}_P\left[\mathbf{c}_P\mathbf{c}_P^\mathsf{T}\right]/d_P\right] + \alpha^2\sigma_P^2 \\
&= (1-\alpha)^2 + \alpha^2\sigma_P^2,
\end{aligned}$$

where we have used the assumptions that $\mathbb{E}_P\left[\mathbf{c}_P\mathbf{c}_P^\mathsf{T}\right] = \mathbf{I}$ and that $\mathbb{E}_P\left[\mathbf{z}\mathbf{z}^\mathsf{T}\right] = \sigma_P^2\mathbf{I}$ again. Similarly, on distribution Q,

$$\begin{aligned}
\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) &= \mathbb{E}_Q\left[\left\|(\mathbf{U}_P\mathbf{c}_Q-\widehat{\mathbf{x}}_\lambda)/\sqrt{d_Q}\right\|_2^2\right] \\
&= \mathbb{E}_Q\left[\left\|\left(\mathbf{I}-\alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\right)\mathbf{U}_Q\mathbf{c}_Q/\sqrt{d_Q}\right\|_2^2\right] + \mathbb{E}_Q\left[\left\|\alpha\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{z}/\sqrt{d_Q}\right\|_2^2\right],
\end{aligned}$$

where the second term in the line above can be readily found to be $\alpha^2\sigma_Q^2 d_P/d_Q$, and the first term can be computed as

$$\begin{aligned}
\mathbb{E}_Q\left[\left\|\left(\mathbf{I}-\alpha\mathbf{U}_P\mathbf{U}_P^\intercal\right)\mathbf{U}_Q\mathbf{c}_Q/\sqrt{d_Q}\right\|_2^2\right] &= \operatorname{tr}\left[\mathbf{U}_Q^\intercal\left(\mathbf{I}+\left(\alpha^2-2\alpha\right)\mathbf{U}_P\mathbf{U}_P^\intercal\right)\mathbf{U}_Q\mathbb{E}_Q\left[\mathbf{c}_Q\mathbf{c}_Q^\intercal\right]/d_Q\right]\\
&= \operatorname{tr}\left[\left(\mathbf{I}+\left(\alpha^2-2\alpha\right)\mathbf{U}_P\mathbf{U}_P^\intercal\right)\mathbf{U}_Q\mathbf{U}_Q^\intercal/d_Q\right]\\
&= \operatorname{tr}\left[\mathbf{U}_Q\mathbf{U}_Q^\intercal/d_Q\right]+\left(\alpha^2-2\alpha\right)\operatorname{tr}\left[\mathbf{U}^\intercal\mathbf{U}_Q\left(\mathbf{U}_P^\intercal\mathbf{U}_Q\right)^\intercal/d_Q\right]\\
&\overset{(c)}{=} 1+\left(\alpha^2-2\alpha\right)\frac{1}{d_Q}\sum_{i=1}^{\min\{d_P,d_Q\}} cos^2(\theta_i)\\
&= 1+(\alpha^2-2\alpha)\frac{\|cos(\boldsymbol{\theta})\|_2^2}{d_Q},
\end{aligned}$$

where (c) follows from the fact that the singular values of $\mathbf{U}_P^\intercal\mathbf{U}_Q$ are the cosines of the principal angle $\theta_i, i\in[\min\{d_P,d_Q\}]$ between $\mathbf{U}_P$ and $\mathbf{U}_Q$. Hence,

$$\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) = 1+(\alpha^2-2\alpha)\frac{\|cos(\boldsymbol{\theta})\|_2^2}{d_Q}+\alpha^2\sigma_Q^2\frac{d_P}{d_Q}.$$

The expression of $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ implies that

$$\alpha^2-2\alpha = \mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)-1-\alpha^2\sigma_P^2.$$

Plugging this expression into the expression of $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ yields the result. ■

E.2 Proof of Theorem 5

We first provide two lemmas which are used in the main proof. In the first lemma, the risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ are expressed in terms of matrices $\mathbf{U}_P^\intercal\mathbf{U}_P$ and $\mathbf{U}_P^\intercal\mathbf{U}_Q$, and their approximations $\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P$ and $\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_Q$ induced by the random measurement matrix $\mathbf{A}$.

Lemma 16 *The risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ of $\widehat{\mathbf{x}}_\lambda$ can be expressed as*

$$\begin{aligned}
\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) &= \left(\left\|\mathbf{I}-\mathbf{S}\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P\right\|_F^2+\sigma_P^2\operatorname{tr}\left(\mathbf{S}^\intercal\mathbf{S}\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P\right)\right)/d_P,\\
\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) &= \left(\left\|\mathbf{U}_P^\intercal\mathbf{U}_Q-\mathbf{S}\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_Q\right\|_F^2-\left\|\mathbf{U}_P^\intercal\mathbf{U}_Q\right\|_F^2+\|\mathbf{U}_Q\|_F^2+\sigma_Q^2\operatorname{tr}\left(\mathbf{S}^\intercal\mathbf{S}\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P\right)\right)/d_Q,
\end{aligned}$$

where $\mathbf{S}=\eta\mathbf{I}-\eta^2\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P\left(\mathbf{I}+\eta\mathbf{U}_P^\intercal\mathbf{A}^\intercal\mathbf{A}\mathbf{U}_P\right)^{-1}$ and $\eta=1/(\sigma_P^2+\lambda)$.

Proof Similarly to the proof of Theorem 4, the solution $\mathbf{W}^*$ to $\min_{\mathbf{W}} \mathbb{E}_P \left[\|\mathbf{x} - \mathbf{W}\mathbf{y}\|_2^2 \right] + \lambda \|\mathbf{W}\|_F^2$ can be computed as

$$\begin{aligned}
\mathbf{W}^* &= \mathbb{E}_P \left[\mathbf{x}\mathbf{y}^\mathsf{T} \right] \left(\mathbb{E}_P \left[\mathbf{y}\mathbf{y}^\mathsf{T} \right] + \lambda \mathbf{I} \right)^{-1} \\
&= \mathbb{E}_P \left[\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T} \right] \left(\mathbb{E}_P \left[\mathbf{A}\mathbf{x}\mathbf{x}^\mathsf{T}\mathbf{A}^\mathsf{T} + \mathbf{z}\mathbf{z}^\mathsf{T} \right] + \lambda \mathbf{I} \right)^{-1} \\
&= \mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T} \left(\mathbf{A}\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T} + (\sigma_P^2 + \lambda)\mathbf{I} \right)^{-1} \\
&\overset{(a)}{=} \mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T} \left(\eta \mathbf{I} - \eta^2 \mathbf{A}\mathbf{U}_P \left(\mathbf{I} + \eta \mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P \right)^{-1} \mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T} \right) \\
&= \mathbf{U}_P\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T},
\end{aligned}$$

where (a) follows from the matrix inversion lemma and $\eta = 1/(\sigma_P^2 + \lambda)$. The risk $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ can be computed as

$$\begin{aligned}
\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) &= \mathbb{E}_P \left[\left\| (\mathbf{U}_P\mathbf{c}_P - \mathbf{W}^*(\mathbf{A}\mathbf{U}_P\mathbf{c}_P + \mathbf{z}))/\sqrt{d_P} \right\|_2^2 \right] \\
&= \mathbb{E}_P \left[\left\| (\mathbf{I} - \mathbf{W}^*\mathbf{A})\,\mathbf{U}_P\mathbf{c}_P/\sqrt{d_P} \right\|_2^2 \right] + \mathbb{E}_P \left[\left\| \mathbf{W}^*\mathbf{z}/\sqrt{d_P} \right\|_2^2 \right] \\
&= \left(\mathrm{tr}\left(\mathbf{U}_P^\mathsf{T}(\mathbf{I} - \mathbf{W}^*\mathbf{A})^\mathsf{T}(\mathbf{I} - \mathbf{W}^*\mathbf{A})\mathbf{U}_P\mathbb{E}_P\left[\mathbf{c}_P\mathbf{c}_P^\mathsf{T} \right] \right) + \mathrm{tr}\left(\mathbf{W}^{*\mathsf{T}}\mathbf{W}\mathbb{E}_P\left[\mathbf{z}\mathbf{z}^\mathsf{T} \right] \right) \right)/d_P \\
&= \left(\|(\mathbf{I} - \mathbf{W}^*\mathbf{A})\mathbf{U}_P\|_F^2 + \sigma_P^2 \mathrm{tr}\left(\mathbf{W}^{*\mathsf{T}}\mathbf{W}^* \right) \right)/d_P \\
&= \left(\left\| \mathbf{U}_P(\mathbf{I} - \mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P) \right\|_F^2 + \sigma_P^2 \mathrm{tr}\left(\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P \right) \right)/d_P \\
&= \left(\left\| \mathbf{I} - \mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P \right\|_F^2 + \sigma_P^2 \mathrm{tr}\left(\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P \right) \right)/d_P.
\end{aligned}$$

Similarly, the risk $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ can be computed as

$$\begin{aligned}
\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) &= \mathbb{E}_Q \left[\left\| (\mathbf{U}_Q\mathbf{c}_Q - \mathbf{W}^*(\mathbf{A}\mathbf{U}_Q\mathbf{c}_Q + \mathbf{z}))/\sqrt{d_Q} \right\|_2^2 \right] \\
&= \mathbb{E}_Q \left[\left\| (\mathbf{I} - \mathbf{W}^*\mathbf{A})\,\mathbf{U}_Q\mathbf{c}_Q/\sqrt{d_Q} \right\|_2^2 \right] + \mathbb{E}_Q \left[\left\| \mathbf{W}^*\mathbf{z}/\sqrt{d_Q} \right\|_2^2 \right],
\end{aligned}$$

where the second term in the line above can be readily found to be $\sigma_Q^2 \mathrm{tr}\left(\mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P\right)/d_Q$, and the first term can be computed as

$$
\begin{aligned}
&\mathbb{E}_Q\left[\left\|(\mathbf{I}-\mathbf{W}^*\mathbf{A})\,\mathbf{U}_Q\mathbf{c}_Q/\sqrt{d_Q}\right\|_2^2\right] \\
&= \mathrm{tr}\left(\mathbf{U}_Q^\mathsf{T}(\mathbf{I}-\mathbf{W}^*\mathbf{A})^\mathsf{T}(\mathbf{I}-\mathbf{W}^*\mathbf{A})\mathbf{U}_Q\mathbb{E}_Q\left[\mathbf{c}_Q\mathbf{c}_Q^\mathsf{T}\right]\right)/d_Q \\
&= \|(\mathbf{I}-\mathbf{W}^*\mathbf{A})\mathbf{U}_Q\|_F^2/d_Q \\
&= \left\|\mathbf{U}_Q-\mathbf{U}_P\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q\right\|_F^2/d_Q \\
&= \left\|(\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q-\mathbf{U}_P\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q)+(\mathbf{I}-\mathbf{U}_P\mathbf{U}_P^\mathsf{T})\mathbf{U}_Q\right\|_F^2/d_Q \\
&\overset{(b)}{=} \left(\left\|\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q-\mathbf{U}_P\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q\right\|_F^2+\left\|(\mathbf{I}-\mathbf{U}_P\mathbf{U}_P^\mathsf{T})\mathbf{U}_Q\right\|_F^2\right)/d_Q \\
&= \left(\left\|\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q-\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q\right\|_F^2+\left\|\mathbf{U}_Q-\mathbf{U}_P\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q\right\|_F^2\right)/d_Q \\
&= \left(\left\|\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q-\mathbf{S}\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q\right\|_F^2-\left\|\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q\right\|_F^2+\|\mathbf{U}_Q\|_F^2\right)/d_Q,
\end{aligned}
$$

where (b) follows from the fact that matrices $\mathbf{U}_P$ and $\mathbf{I}-\mathbf{U}_P\mathbf{U}_P^\mathsf{T}$ are orthogonal under the Frobenius inner product. ■

The second lemma below about inner product preservation is used to show that matrices $\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_P$ and $\mathbf{U}_P^\mathsf{T}\mathbf{A}^\mathsf{T}\mathbf{A}\mathbf{U}_Q$ are close to $\mathbf{U}_P^\mathsf{T}\mathbf{U}_P$ and $\mathbf{U}_P^\mathsf{T}\mathbf{U}_Q$ element-wise respectively.

Lemma 17 *Let $\mathbf{A} \in \mathbb{R}^{n\times d}$ be a random Gaussian matrix with independent entries drawn from the distribution $\mathcal{N}(0, 1/n)$. For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ with $\|\mathbf{u}\|_2 \le 1, \|\mathbf{v}\|_2 \le 1$ and $0 < \epsilon < 1$, with probability at least $1 - 4\exp(-n\epsilon^2/8)$,*

$$|\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle| \le \epsilon.$$

Proof The proof relies on the result on norm preservation by random projection: for any $0 < t < 1$, it holds that

$$\Pr\left(\left|\frac{\|\mathbf{A}\mathbf{u}\|_2^2}{\|\mathbf{u}\|_2^2} - 1\right| \ge t\right) \le 2e^{-\frac{nt^2}{8}},$$

which follows from the fact that, for any $\mathbf{u} \in \mathbb{R}^d$, the variable $\|\mathbf{A}\mathbf{u}\|_2^2/\|\mathbf{u}\|_2^2$ follows the same distribution as $(1/n)\chi^2(n)$ and that a $(1/n)\chi^2(n)$ distribution is sub-exponential with parameters $(2^2/n, 4)$.

Now consider any $\mathbf{u}, \mathbf{v}$ with $\|\mathbf{u}\|_2 \le 1, \|\mathbf{v}\|_2 \le 1$ and $0 < \epsilon < 1$. Using the fact that $\langle \mathbf{u}, \mathbf{v}\rangle = (1/4)\left(\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2\right)$, under the events that the norm squares $\|\mathbf{u}+\mathbf{v}\|_2^2$

and $\|\mathbf{u}-\mathbf{v}\|_2^2$ are approximately preserved, which occur with probability at least $1 - 4\exp(-n\epsilon^2/8)$, it holds that

$$\begin{aligned}\langle \mathbf{Au}, \mathbf{Av}\rangle &= \frac{1}{4}\Big(\|\mathbf{u}+\mathbf{v}\|_2^2 - \|\mathbf{u}-\mathbf{v}\|_2^2\Big) \\ &\le \frac{1}{4}\Big((1+\epsilon)\|\mathbf{u}+\mathbf{v}\|_2^2 - (1-\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2\Big) \\ &= \frac{1}{4}\Big(4\langle \mathbf{u}, \mathbf{v}\rangle + 2\epsilon\|\mathbf{u}\|_2^2 + 2\epsilon\|\mathbf{v}\|_2^2\Big) \\ &\le \langle \mathbf{u}, \mathbf{v}\rangle + \epsilon,\end{aligned}$$

and that

$$\langle \mathbf{Au}, \mathbf{Av}\rangle \ge \langle \mathbf{u}, \mathbf{v}\rangle - \epsilon,$$

following a similar derivation. ■

E.2.1 Proof of Theorem 5

The proof idea is essentially the same as the proof of Theorem 4: expressing the risks of the estimate $\widehat{\mathbf{x}}_\lambda$ on distributions P and Q as functions of $\alpha = 1/(1+\sigma_P^2+\lambda)$ and then expressing the risk $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ in terms of $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$. The only technical issue is that $\mathbf{W}^*\mathbf{y}$ is only approximately $\alpha\mathbf{U}_P\mathbf{U}_P^\top\mathbf{y}$ due to the random measurement by matrix $\mathbf{A}$. We show that, under the event that the map $\mathbf{u} \mapsto \mathbf{Au}$ approximately preserves inner products of interest, as defined below, the risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ can be expressed respectively as those in the proof of Theorem 4 plus some error terms which converge to zero as the number of measurements $n \to \infty$ with high probability.

Step 1. Expressing matrices $\mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_P$, $\mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_Q$ and $\mathbf{S}$ as perturbed matrices. For any $0 < \epsilon < 1/d_P$, consider the event

$$\mathcal{E}:\ |\langle \mathbf{Au}, \mathbf{Av}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle| \le \epsilon, \quad \begin{aligned}&\text{for all pairs of columns } (\mathbf{u}, \mathbf{v}) \text{ of } \mathbf{U}_P \text{ and} \\ &\text{for all column } \mathbf{u} \text{ of } \mathbf{U}_P \text{ and column } \mathbf{v} \text{ of } \mathbf{U}_Q,\end{aligned}$$

which happens with probability at least $1 - 4(d_P^2 + d_P d_Q)\exp(-n\epsilon^2/8)$ by Lemma 17 and the union bound. Under event $\mathcal{E}$, matrices $\mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_P$ and $\mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_Q$ are perturbed versions of $\mathbf{I}$ and $\mathbf{U}_P^\top\mathbf{U}_Q$, i.e.,

$$\begin{aligned}\mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_P &= \mathbf{I} + \mathbf{E}, \quad |\mathbf{E}_{ij}| \le \epsilon, \quad \forall\, i \in [d_P], \quad \forall\, j \in [d_P], \\ \mathbf{U}_P^\top\mathbf{A}^\top\mathbf{A}\mathbf{U}_Q &= \mathbf{U}_P^\top\mathbf{U}_Q + \mathbf{F}, \quad |\mathbf{F}_{ij}| \le \epsilon, \quad \forall\, i \in [d_P], \quad \forall\, j \in [d_Q],\end{aligned}$$

and, as a result, $\mathbf{S}$ is a pertubed version of $\alpha\mathbf{I}$, i.e.,

$$\mathbf{S} = \alpha\mathbf{I} + \mathbf{G},$$

where matrix $\mathbf{G}$ is a polynomial of matrix $\mathbf{E}$ as defined below, since

$$\begin{aligned}
\mathbf{S} &= \eta\mathbf{I} - \eta^2(\mathbf{I}+\mathbf{E})\left(\mathbf{I}+\eta(\mathbf{I}+\mathbf{E})\right)^{-1}\\
&= \eta\mathbf{I} - \frac{\eta^2}{1+\eta}(\mathbf{I}+\mathbf{E})\left(\mathbf{I}+\frac{\eta}{1+\eta}\mathbf{E}\right)^{-1}\\
&\overset{(a)}{=} \eta\mathbf{I} - \frac{\eta^2}{1+\eta}(\mathbf{I}+\mathbf{E})\sum_{k=0}^{\infty}\left(-\frac{\eta}{1+\eta}\mathbf{E}\right)^k\\
&= \eta\mathbf{I} - \frac{\eta^2}{1+\eta}\left(\mathbf{I}+\mathbf{E}+(\mathbf{I}+\mathbf{E})\sum_{k=1}^{\infty}\left(-\frac{\eta}{1+\eta}\mathbf{E}\right)^k\right)\\
&= \left(\eta-\frac{\eta^2}{1+\eta}\right)\mathbf{I} - \frac{\eta^2}{1+\eta}\left(\mathbf{E}+(\mathbf{I}+\mathbf{E})\sum_{k=1}^{\infty}\left(-\frac{\eta}{1+\eta}\mathbf{E}\right)^k\right)\\
&= \alpha\mathbf{I} - \frac{\alpha^2}{1-\alpha}\left(\mathbf{E}+(\mathbf{I}+\mathbf{E})\sum_{k=1}^{\infty}(-\alpha\mathbf{E})^k\right),
\end{aligned}$$

with $\mathbf{G} = -(\alpha^2/(1-\alpha))\left(\mathbf{E}+(\mathbf{I}+\mathbf{E})\sum_{k=1}^{\infty}(-\alpha\mathbf{E})^k\right)$ and $\alpha = 1/(1+\sigma_P^2+\lambda)$. Recall that $\eta = 1/(\sigma_P^2+\lambda)$. Step (a) is valid because $(\eta/(1+\eta))\mathbf{E}$ has eigenvalues bounded by 1, since for $\epsilon < 1/d_P$, $\max_i |\lambda_i((\eta/(1+\eta))\mathbf{E})| \le |\lambda_i(\mathbf{E})| \le (\sum_i \lambda_i^2(\mathbf{E}))^{1/2} = \|\mathbf{E}\|_F \le \epsilon d_P \le 1$.

Step 2. Expressing risks $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ and $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ as functions of α and the perturbation matrices.
Under event $\mathcal{E}$, it holds that

$$\begin{aligned}
\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda) &= \|\mathbf{I}-(\alpha\mathbf{I}+\mathbf{G})(\mathbf{I}+\mathbf{E})\|_F^2/d_P + \operatorname{tr}\left((\alpha\mathbf{I}+\mathbf{G})^2(\mathbf{I}+\mathbf{E})\right)\sigma_P^2/d_P\\
&= \|(1-\alpha)\mathbf{I}-((\alpha\mathbf{I}+\mathbf{G})\mathbf{E}+\mathbf{G})\|_F^2/d_P + \operatorname{tr}\left(\alpha^2\mathbf{I}+\alpha^2\mathbf{E}+(2\alpha\mathbf{G}+\mathbf{G}^2)(\mathbf{I}+\mathbf{E})\right)\sigma_P^2/d_P\\
&= (1-\alpha)^2 + \epsilon_{\text{signal, P}} + \alpha^2\sigma_P^2 + \epsilon_{\text{noise, P}},
\end{aligned}$$

where $\epsilon_{\text{signal, P}}$ and $\epsilon_{\text{noise, P}}$ are errors induced by the random measurement and are expressed as

$$\begin{aligned}
\epsilon_{\text{signal, P}} &= \operatorname{tr}\big([(\alpha\mathbf{I}+\mathbf{G})\mathbf{E}+\mathbf{G}-2(1-\alpha)\mathbf{I}][(\alpha\mathbf{I}+\mathbf{G})\mathbf{E}+\mathbf{G}]\big)/d_P,\\
\epsilon_{\text{noise, P}} &= \operatorname{tr}\left(\alpha^2\mathbf{E}+(2\alpha\mathbf{G}+\mathbf{G}^2)(\mathbf{I}+\mathbf{E})\right)\sigma_P^2/d_P,
\end{aligned}$$

and that

$$\begin{aligned}
\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda) &= \left(\left\|\mathbf{U}_P^\intercal\mathbf{U}_Q-(\alpha\mathbf{I}+\mathbf{G})(\mathbf{U}_P^\intercal\mathbf{U}_Q+\mathbf{F})\right\|_F^2 - \left\|\mathbf{U}_P^\intercal\mathbf{U}_Q\right\|_F^2 + \|\mathbf{U}_Q\|_F^2\right)/d_Q\\
&\qquad + \operatorname{tr}\left((\alpha\mathbf{I}+\mathbf{G})^2(\mathbf{I}+\mathbf{E})\right)\sigma_Q^2/d_Q\\
&= \left(\left\|(1-\alpha)\mathbf{U}_P^\intercal\mathbf{U}_Q-((\alpha\mathbf{I}+\mathbf{G})\mathbf{F}+\mathbf{G}\mathbf{U}_P^\intercal\mathbf{U}_Q)\right\|_F^2 - \left\|\mathbf{U}_P^\intercal\mathbf{U}_Q\right\|_F^2 + \|\mathbf{U}_Q\|_F^2\right)/d_Q\\
&\qquad + \operatorname{tr}\left(\alpha^2\mathbf{I}+\alpha^2\mathbf{E}+(2\alpha\mathbf{G}+\mathbf{G}^2)(\mathbf{I}+\mathbf{E})\right)\sigma_Q^2/d_Q\\
&= (\alpha^2-2\alpha)\frac{\left\|\mathbf{U}_P^\intercal\mathbf{U}_Q\right\|_F^2}{d_Q} + 1 + \epsilon_{\text{signal, Q}} + \alpha^2\sigma_Q^2\frac{d_P}{d_Q} + \epsilon_{\text{noise, Q}},
\end{aligned}$$

where $\epsilon_{\text{signal, Q}}$ and $\epsilon_{\text{noise, Q}}$ are also errors induced by the random measurement and are expressed as

$$\begin{aligned}\epsilon_{\text{signal, Q}} &= \operatorname{tr}\big([(\alpha\mathbf{I}+\mathbf{G})\mathbf{F}+\mathbf{G}\mathbf{U}_P^\top\mathbf{U}_Q-2(1-\alpha)\mathbf{U}_P^\top\mathbf{U}_Q]^\top[(\alpha\mathbf{I}+\mathbf{G})\mathbf{F}+\mathbf{G}\mathbf{U}_P^\top\mathbf{U}_Q]\big)/d_Q,\\ \epsilon_{\text{noise, Q}} &= \frac{\sigma_Q^2}{\sigma_P^2}\frac{d_P}{d_Q}\epsilon_{\text{noise, P}}.\end{aligned}$$

The expression of $\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)$ implies that

$$\alpha^2-2\alpha=\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)-1-\alpha^2\sigma_P^2-\epsilon_{\text{signal, P}}-\epsilon_{\text{noise, P}}.$$

Plugging this expression into the expression of $\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)$ yields

$$\mathcal{R}_Q(\widehat{\mathbf{x}}_\lambda)=a\mathcal{R}_P(\widehat{\mathbf{x}}_\lambda)+(1-a)+\alpha^2\left(\frac{d_P}{d_Q}\sigma_Q^2-a\sigma_P^2\right)+\epsilon_{\text{signal, Q}}+\epsilon_{\text{noise, Q}}-a(\epsilon_{\text{signal, P}}+\epsilon_{\text{noise, P}}),$$

where $a=\left\|\mathbf{U}_P^\top\mathbf{U}_Q\right\|_F^2/d_Q=\|cos(\boldsymbol{\theta})\|_2^2/d_Q$ and $\boldsymbol{\theta}\in\mathbb{R}^{\min\{d_P,d_Q\}}$ is the principal angles between $\mathbf{U}_P$ and $\mathbf{U}_Q$.

Step 3. Bounding the errors caused by the perturbation matrices.
It remains to show that, under event $\mathcal{E}$, the error term $\epsilon_{\text{signal, Q}}+\epsilon_{\text{noise, Q}}-a(\epsilon_{\text{signal, P}}+\epsilon_{\text{noise, P}})$ is bounded by some constant times ϵ. We show that each error in the error term is $O(\epsilon)$.

With some computation, it is easy to check that each of the errors $\epsilon_{\text{signal, P}}$, $\epsilon_{\text{noise, P}}$ and $\epsilon_{\text{signal, Q}}$ is the trace of a polynomial of the perturbation matrices, i.e., there exist polynomials $p_1,\dots,p_5$ such that

$$\begin{aligned}\epsilon_{\text{signal, P}} &= \operatorname{tr}(p_1(\mathbf{E})),\\ \epsilon_{\text{noise, P}} &= \operatorname{tr}(p_2(\mathbf{E})),\\ \epsilon_{\text{signal, Q}} &= \operatorname{tr}(p_3(\mathbf{E})\mathbf{F}\mathbf{F}^\top)+\operatorname{tr}(p_4(\mathbf{E})\mathbf{F}\mathbf{U}_Q^\top\mathbf{U}_P)+\operatorname{tr}(p_5(\mathbf{E})\mathbf{U}_P^\top\mathbf{U}_Q\mathbf{U}_Q^\top\mathbf{U}_P),\end{aligned}$$

and that p_1, p_2 and p_5 have zero-th order terms zeros. Recall that matrices $\mathbf{E}$ and $\mathbf{F}$ have entries bounded by ϵ. Therefore, it holds that $\|\mathbf{E}\|_F\le\epsilon d_P$ and $\|\mathbf{F}\|_F\le\epsilon\sqrt{d_Pd_Q}$, and that

$$|\operatorname{tr}(\mathbf{E}^k)|=\big|\sum_i\sigma_i^k(\mathbf{E})\big|\overset{(b)}{\le}\big|\sum_i\sigma_i(\mathbf{E})\big|\le\epsilon d_P,\quad\forall k\ge1,$$

where (b) follows from the fact that $\max_i|\sigma_i(\mathbf{E})|\le\|\mathbf{E}\|_F\le\epsilon d_P$ for $\epsilon<1/d_P$. As a result, $\epsilon_{\text{signal, P}}$ and $\epsilon_{\text{noise, P}}$ are $O(\epsilon)$. So is $\epsilon_{\text{signal, Q}}$, because each of its term is $O(\epsilon)$. Indeed, for any $k\ge0$,

$$\begin{aligned}\operatorname{tr}(\mathbf{E}^k\mathbf{F}\mathbf{F}^\top) &\le \operatorname{tr}^{\frac12}(\mathbf{E}^{2k})\left\|\mathbf{F}\mathbf{F}^\top\right\|_F\le\operatorname{tr}^{\frac12}(\mathbf{I})\,\|\mathbf{F}\|_F^2\le\sqrt{d_P}\,\epsilon^2d_Pd_Q,\\ \operatorname{tr}(\mathbf{E}^k\mathbf{F}\mathbf{U}_Q^\top\mathbf{U}_P) &\le \operatorname{tr}^{\frac12}(\mathbf{E}^{2k})\left\|\mathbf{F}\mathbf{U}_Q^\top\mathbf{U}_P\right\|_F\le\operatorname{tr}^{\frac12}(\mathbf{I})\left\|\mathbf{U}_Q^\top\mathbf{U}_P\right\|_2\|\mathbf{F}\|_F\le\sqrt{d_P}\,\epsilon\sqrt{d_Pd_Q},\end{aligned}$$

and for any $k\ge1$,

$$\begin{aligned}\operatorname{tr}(\mathbf{E}^k\mathbf{U}_P^\top\mathbf{U}_Q\mathbf{U}_Q^\top\mathbf{U}_P)\le\operatorname{tr}^{\frac12}(\mathbf{E}^{2k})\left\|\mathbf{U}_P^\top\mathbf{U}_Q\mathbf{U}_Q^\top\mathbf{U}_P\right\|_F &\le\operatorname{tr}^{\frac12}(\mathbf{E}^2)\left\|\mathbf{U}_P^\top\mathbf{U}_Q\right\|_F^2\\ &\overset{(c)}{\le}\epsilon d_P\min\{d_P,d_Q\},\end{aligned}$$

where (c) follows from the fact that $\left\|\mathbf{U}_P^\intercal \mathbf{U}_Q\right\|_F^2 = \|cos(\boldsymbol{\theta})\|_2^2$. We conclude that the error term $\epsilon_{\text{signal, Q}} + \epsilon_{\text{noise, Q}} - a(\epsilon_{\text{signal, P}} + \epsilon_{\text{noise, P}}) = O(\epsilon)$ and the proof is complete. ■

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(12):151175, 2010.

Benjamin Caine, Rebecca Roelofs, Vijay Vasudevan, Jiquan Ngiam, Yuning Chai, Zhifeng Chen, and Jonathon Shlens. Pseudo-labeling for scalable 3D object detection. *arXiv preprint arXiv:2103.02093*, 2021.

Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

Mohammad Zalbagi Darestani, Akshay S Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *International Conference on Machine Learning*, volume 139, pages 2433–2444, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, volume 119, pages 2892–2901, 2020.

Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303338, 2010.

Sebastian Goldt, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10:041044, Dec 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, 2017.

Reinhard Heckel and Helmut Bölcskei. Robust subspace clustering via thresholding. *IEEE Transactions on Information Theory*, 61(11):6320–6342, 2015.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, Laughing, tkianai, Adam Hogan, lorenzomammana, yxNONG, AlexWang1900, Laurentiu Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Francisco Ingham, Frederik, Guilhen, Hatovix, Jake Poznanski, Jiacong Fang, Lijun Yu, changyu98, Mingyu Wang, Naman Gupta, Osama Akhtar, PetrDvoracek, and Prashant Rai. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements, 2020. URL `https://doi.org/10.5281/zenodo.4154370`.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Huseyin Kusetogullari, Amir Yavariabdi, Abbas Cheddad, Håkan Grahn, and Johan Hall. ARDIS: A Swedish historical handwritten digit dataset. *Neural Computing and Applications*, 32(21):1650516518, 2020.

Yann LeCun, Corinna Cortes, and CJ Burges. MNIST handwritten digit database, 2010. URL `http://yann.lecun.com/exdb/mnist`.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision*, 2017.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, volume 34, pages 18137–18151, 2021.

Horia Mania and Suvrit Sra. Why do classifier accuracies show linear trends under distribution shift? *arXiv preprint arXiv:2012.15483*, 2020.

R. Thomas McCoy, Junghyun Min, and Tal Linzen. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 217–227, 2020.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. In *International Conference on Machine Learning*, volume 119, pages 6905–6916, 2020.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. In *International Conference on Machine Learning*, volume 139, pages 7721–7735, 2021.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset shift in machine learning.* MIT Press, 2008.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, volume 97, pages 5389–5400, 2019.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations*, 2015.

Mahdi Soltanolkotabi and Emmanuel J. Candès. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64 (8):5592–5628, 2018.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

Chhavi Yadav and Léon Bottou. Cold case: The lost MNIST digits. In *Advances in Neural Information Processing Systems*, volume 32, 2019.