Monotonic Risk Relationships under Distribution Shifts for Regularized Risk Minimization

Daniel LeJeune [*1] **Jiayu Liu** [*2] **Reinhard Heckel** [1 2]

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 work, 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 under a subspace shift model with feature scaling.

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 in-distribution and out-of-distribution performance relate.

While there are settings in which models with similar in-distribution performance have different out-of-distribution performance (McCoy et al., 2019), several recent empirical studies have shown that often, the in-distribution and out-of-distribution performances of models are strongly correlated.

Recht et al. (2018; 2019), Yadav & Bottou (2019), 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 all four cases, the performance drops significantly when evaluated on the new test set, indicating that even when following the original data collection and labeling process, a distribution shift occurs. 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 persists for image reconstruction methods that are only tuned and not trained.

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 for a class of distribution shifts where the features come from different distributions.

We consider a general setup encompassing classification and regression for a large class of estimators obtained with

[*]Equal contribution [1]Department of Electrical and Computer Engineering, Rice University [2]Department of Electrical and Computer Engineering, Technical University of Munich. Correspondence to: Daniel LeJeune <daniel@dlej.net>, Reinhard Heckel <reinhard.heckel@gmail.com>.

Presented at ICML Workshop on Principles of Distribution Shift 2022,

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 of the estimators is monotonic. Different classes of estimators using different feature spaces follow different monotonic relations, and we also observe this in practice (see Figure 2). 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 by Miller et al. (2021).

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.

Prior work. Classical theory for characterizing out-of-distribution performance ensures that the difference between in- and out-of-distribution performances 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 & Mohri, 2014). Such bounds often apply to a class of target distributions. In contrast, we are interested in precise relationships between two fixed distributions.

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 & 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 correctly classified by lower accuracy models, providing a different approach to explaining the linear relation phenomenon without characterizing the distribution shift.

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 subspace shift we propose, 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 also add a more general task-dependent feature scaling, which captures how classification problems can become easier or harder. Moreover, our results predict general monotonic relationships as opposed to only linear ones.

2. Main Results

We prove a general monotonic risk relationship for *all models* trained using ridge-regularized empirical risk minimization (ERM) under a subspace shift model in the proportional asymptotics regime, in which the limit of the number of features over the number of training examples converges to a constant, i.e., $\lim_{d\to\infty} d/n \in (0, \infty)$.

We consider data where the function for generating labels from features, which we refer to as the *task*, depends only on a fixed linear combination of features. The setup encompasses both classification and regression. The distribution shift model consists of a subspace shift plus a feature scaling which may correlate with the task. We show that when the shift is task-independent, there is a linear relation between the in- and out-of-distribution squared error. We also show that for more general *task-dependent* shifts, there is a monotonic risk relation for the misclassification error.

Distribution shift model. We consider data drawn from a linear model, where the dependent variable is a function of the inner product of a task vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ and a feature vector $\mathbf{x} \in \mathbb{R}^d$. We assume that the task vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ is obtained by drawing each of its elements independently from a zero-mean random variable B^* with variance σ_β^2, and is then fixed.

We consider two different distributions of the features:

$$P\colon \mathbf{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_P), \quad Q\colon \mathbf{x} = \tilde{\mathbf{x}} \circ \mathbf{s}, \tilde{\mathbf{x}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_Q),$$

where $\boldsymbol{\Sigma}_P$ and $\boldsymbol{\Sigma}_Q$ are covariance matrices defined below, and $\circ$ denotes entry-wise multiplication. The vector $\mathbf{s} \in \mathbb{R}^d$ is a feature scaling of the form $[\mathbf{s}]_j = \sqrt{\rho} s(|[\boldsymbol{\beta}^*]_j|)$ for a feature scaling function $s\colon [0, \infty) \to [0, \infty)$, and $\rho > 0$ is a scaling factor.

Our distribution shift model has two components. The first component is a subspace shift model on the covariance matrices $\boldsymbol{\Sigma}_P, \boldsymbol{\Sigma}_Q$. We partition the feature space $\mathbb{R}^d$ into three subspaces: one unique to distribution P, one unique to distribution Q, and one shared by both, which we label as subspace R. Define diagonal pairwise orthogonal subspace projection matrices $\boldsymbol{\Pi}_P, \boldsymbol{\Pi}_Q, \boldsymbol{\Pi}_R \in \mathbb{R}^{d\times d}$ such that $\boldsymbol{\Pi}_P + \boldsymbol{\Pi}_Q + \boldsymbol{\Pi}_R = \mathbf{I}_d$, and define the covariance matrices as

$$\boldsymbol{\Sigma}_P = \frac{1}{d}(\boldsymbol{\Pi}_P + \boldsymbol{\Pi}_R) \quad \text{and} \quad \boldsymbol{\Sigma}_Q = \frac{1}{d}(\boldsymbol{\Pi}_Q + \boldsymbol{\Pi}_R).$$

The second component of the distribution shift is a *task-dependent* feature scaling imposed by entry-wise multiplication with the scaling vector $\mathbf{s}$. The scaling changes the task difficulty by changing the value placed on features that are easier or harder to learn from the training distribution.

We assume the scaling function s to be normalized such that $\mathbb{E}\left[s(B^*)^2 B^{*2}\right] = \mathbb{E}\left[B^{*2}\right]$, since we already control the

overall scale with the scaling parameter ρ. As we show, this means that the effect of the choice of scaling function s is determined entirely by its second moment, which we call $\varsigma^2 \triangleq \mathbb{E}\left[s(B^*)^2\right]$. Changes in the scaling function s result in an increase or decrease of the error of our estimate $\widehat{\boldsymbol{\beta}}$ by a factor of ς in a way that is independent of the subspace shift. If $\varsigma > 1$, then s is a decreasing function of its input and places more emphasis on features that were not utilized as much in training, making the problem more difficult. On the other hand, when $\varsigma < 1$, the features are scaled increasingly in their magnitude, and the labels rely primarily on features that would have been learned well during training, making it easier. If $\varsigma = 1$, corresponding to $s(b) = 1$, then the distribution shift is task-independent. A simple choice of s is $s(b) = |b|^p \sigma_\beta / \sqrt{\mathbb{E}\left[|B^*|^{2(p+1)}\right]}$, where $p \geq 0$ results in $\varsigma \leq 1$ and $p < 0$ gives $\varsigma > 1$. For Gaussian B^*, this choice of s gives the simple form $\varsigma = 1/\sqrt{2p+1}$, which can range from 0 to ∞ for $p \in (-1/2, \infty)$.

We are interested in characterizing the relationship between the risks of an estimate $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}^*$ defined with respect to an error metric $\psi\colon \mathbb{R}^2 \to \mathbb{R}$ on distributions P and Q:

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

We consider the squared error $\psi(z^*, \hat{z}) = (z^* - \hat{z})^2$ and the misclassification error $\psi(z^*, \hat{z}) = \mathbb{1}\{z^* \hat{z} < 0\}$ as error metrics for regression and classification, respectively.

Ridge-regularized empirical risk minimization (ERM). We are given a training dataset $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$ for the task $\boldsymbol{\beta}^*$ generated by drawing the pairs $(\mathbf{x}_i, y_i)$ i.i.d. as

$$\mathbf{x}_i \sim P, \quad y_i = \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}^*, \xi_i), \quad \xi_i \sim \mathcal{N}(0,1).$$

Here $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ is a *labeling function*. We assume the labeling function to be either Lipschitz continuous, such as $\varphi(z, \xi) = z + \sigma\xi$, which results in a linear regression model with Gaussian noise $\mathcal{N}(0, \sigma^2)$, or a bounded function continuous almost everywhere, such as the binary labeling functions $\phi(z, \xi) \sim \text{Bernoulli}(f(z))$ for some $f\colon \mathbb{R} \to [0,1]$. The latter includes common classification labeling schemes such as the logistic model and constant label corruption probability.

We construct our estimate by solving the ridge-regularized ERM formulation for some $\lambda \geq \lambda^* > 0$:

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

Here we assume the loss function $\ell\colon \mathbb{R}^2 \to \mathbb{R}$ is a twice-differentiable proper, closed, convex function satisfying

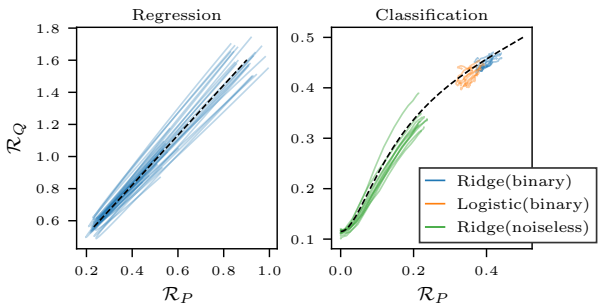

Figure 1: The risk relationships for data generated according to our distribution shift model match our theoretical results (dashed). Each colored curve corresponds to a sweep of the regularization strength of a single model on a single random trial. For both plots, we use $n = 1000$, $d = 800$, $\kappa_P = 0.2$, $\kappa_Q = 0.1$, $\kappa_R = 0.7$, $\sigma_\beta^2 = 1$, and $\rho = 2$. **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 $\varsigma^2 = 1$. Although the tuning parameter overshoots the minimizer in the parameter sweep, it still always lies 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 $\Pr(y_i = \text{sign}(\mathbf{x}_i^\top \boldsymbol{\beta}^*)) = 0.8$ with $\varsigma^2 = 5$. 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.

$|z'| \leq C(1 + |z|)$ for any $z \in \text{dom}(\ell(y, \cdot))$ and $z' \in \partial\ell(y, z)$ for some universal $C \geq 0$. This is satisfied by most common losses used for ERM, such as the squared loss, the logistic loss, and the robust Huber loss. We also are limited to regularization strengths larger than some $\lambda^* > 0$, which is a technical detail; as described by Gerbelot et al. (2020), it is possible to extend to all $\lambda > 0$ for a restricted class of losses and evaluation metrics.

We are now ready to state our result, which holds in the asymptotic regime, where the dimensions d_P, d_Q, and d_R of the three subspaces ($d_P + d_Q + d_R = d$) converge as $d_P/d \to \kappa_P$, $d_Q/d \to \kappa_Q$, and $d_R/d \to \kappa_R$ when $d \to \infty$. Our result leverages recent results from the approximate message passing (AMP) literature (Gerbelot et al., 2020), but could also be proven using the convex Gaussian min-max theorem (Thrampoulidis et al., 2018); see Appendix B.2.

Theorem 2.1. *In the limit as $d \to \infty$ with $\lim_{d\to\infty} d/n \in (0, \infty)$, for any $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathcal{D}, \ell, \lambda)$ solving (1), we have the following monotonic relationships between $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}})$ and $\mathcal{R}_P(\widehat{\boldsymbol{\beta}})$ almost surely:*

(a) Regression. For $\psi(z^*, \hat{z}) = (z^* - \hat{z})^2$ and $\varsigma = 1$,

$$\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) = \theta^2 \mathcal{R}_P(\widehat{\boldsymbol{\beta}}) + \rho\kappa_Q\sigma_\beta^2,$$

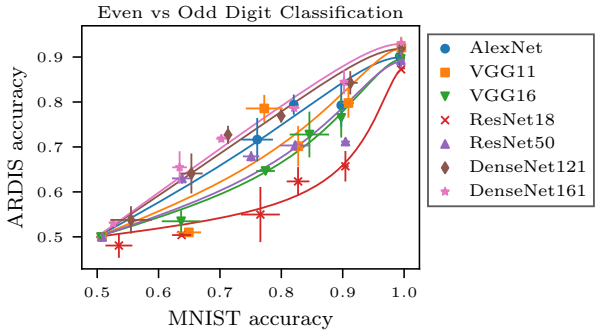

Figure 2: We train deep network models on classifying even vs. odd handwritten digits from the MNIST and ARDIS datasets, evaluating test performance during training as validation accuracy milestones are reached (dots with errorbars over 3 trials). We also plot our theoretical risk relation with κ_Q/κ_R and ς chosen to minimize squared error of the fit for each model.

where $\theta = \sqrt{\rho\kappa_R/(\kappa_R+\kappa_P)}$.

(b) Classification. For $\psi(z^*,\hat{z}) = \mathbb{1}\{z^*\hat{z} < 0\}$,

$$\sec(\pi\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}))^2 = (1+\tfrac{\kappa_Q}{\kappa_R})(1+\varsigma^2\tan(\pi\mathcal{R}_P(\widehat{\boldsymbol{\beta}}))^2).$$

Furthermore, when $\varsigma \geq 1$,

$$\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) > \mathcal{R}_P(\widehat{\boldsymbol{\beta}}),$$

and for any ς*, in the limit as* $\kappa_Q/\kappa_R \to 0^+$,

$$\log(\tan(\pi\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}))) = \log(\tan(\pi\mathcal{R}_P(\widehat{\boldsymbol{\beta}}))) + \log\varsigma.$$

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

Figure 1 illustrates this behavior in finite dimensions; there we plot the prediction of our theory along with instances of our estimates. That is, 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.

Our result also matches the observation by Miller et al. (2021) that the risk relationship for misclassification error as $\kappa_Q/\kappa_R \to 0$ is *linear* with some constant offset after applying a nonlinear transformation. Their choice to use an inverse Gaussian cumulative distribution function transformation $\Phi^{-1}(\cdot)$ is remarkably similar to what we obtain—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" distributions shifts formed by repeated dataset collection may have no subspace shift component ($\kappa_Q/\kappa_R \to 0$), but rather only a task-dependent shift ($\varsigma \neq 1$).

For different feature spaces, our theory predicts different monotonic relations. This is also observed in practice: in Figure 2, we show that except for the VGG11 and ResNet50 models, our theory is very well predictive of the risk relation as a function of early stopping for deep network models trained on MNIST (LeCun et al., 2010), an easy handwritten digits task, and applied to ARDIS (Kusetogullari et al., 2020), a much more difficult handwritten digits dataset. The fits shows that different neural network models, which have their own respective implicit feature spaces, result in different monotonic risk relations.

We remark that the existence of this monotonic relationship doesn't hold for every choice of metric ψ, even for those commonly used for learning. Simple counterexamples are the logistic and hinge losses, as we demonstrate in Appendix A.2. The way in which monotonicity arises even for squared error and misclassification error is quite different between the two cases; see Appendix B.3.

Comments on the proof and possible extensions. Theorem 2.1 is a consequence of the fact that in the asymptotic regime, the estimate converges to the simple form of $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Pi}_P + \boldsymbol{\Pi}_R)(\alpha\boldsymbol{\beta}^* + \mathbf{e})$ for some $\alpha > 0$ and $\mathbf{e}$ that is orthogonal to $\boldsymbol{\beta}^*$, allowing us to very simply characterize the predictions in terms of a few independent Gaussian random variables. The implication of this is that the risk $\mathcal{R}(\widehat{\boldsymbol{\beta}})$ for a general metric ψ for a particular estimate $\widehat{\boldsymbol{\beta}}$ is a function *only* of α and $\|\mathbf{e}\|_2$ for fixed training and test distributions and task $\boldsymbol{\beta}^*$. See Appendix B for proof details.

We have kept our assumptions simple to aid in interpretability, but they have several very straightforward extensions. One is that the data need not be Gaussian, but can come from the more general class of rotationally invariant data distributions with arbitrary spectrum, which is more general than i.i.d. sub-Gaussian data and relevant for random feature models (Pennington & Worah, 2017; Péché, 2019). The assumption that the projection matrices are diagonal and aligned with the standard basis can be relaxed to subspaces in free position if we require B^* to be Gaussian. Lastly, the results extend in the task-independent setting ($\varsigma = 1$) to other convex separable regularizers such as the ℓ_1 penalty, as we describe in Appendix B.4; to capture task-dependent shifts, we state our results only for the ridge penalty.

References

Bayati, M. and Montanari, A. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57

(2):764–785, 2011.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1–2):151–175, 2010.

Caine, B., Roelofs, R., Vasudevan, V., Ngiam, J., Chai, Y., Chen, Z., and Shlens, J. Pseudo-labeling for scalable 3D object detection. *arXiv preprint arXiv:2103.02093*, 2021.

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

Darestani, M. Z., Chaudhari, A. S., and Heckel, R. Measuring robustness in deep learning based compressive sensing. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2433–2444, 2021.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Emami, M., Sahraee-Ardakan, M., Pandit, P., Rangan, S., and Fletcher, A. Generalization error of generalized linear models in high dimensions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 2892–2901, 2020.

Gerbelot, C., Abbara, A., and Krzakala, F. Asymptotic errors for teacher-student convex generalized linear models (or : How to prove Kabashima's replica formula). *ArXiv*, abs/2006.06581, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

Kusetogullari, H., Yavariabdi, A., Cheddad, A., Grahn, H., and Hall, J. ARDIS: A Swedish historical handwritten digit dataset. *Neural Computing and Applications*, 32 (21):16505–16518, 2020.

LeCun, Y., Cortes, C., and Burges, C. MNIST handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist.

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

McCoy, R. T., Min, J., and Linzen, T. BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6905–6916, 2020.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7721–7735, 2021.

Péché, S. A note on the Pennington-Worah distribution. *Electronic Communications in Probability*, 24:1–7, 2019.

Pennington, J. and Worah, P. Nonlinear random matrix theory for deep learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset shift in machine learning*. MIT Press, 2008.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do CIFAR-10 classifiers generalize to CIFAR-10? *arXiv preprint arXiv:1806.00451*, 2018.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5389–5400, 2019.

Salehi, F., Abbasi, E., and Hassibi, B. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

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

Tripuraneni, N., Adlam, B., and Pennington, J. Covariate shift in high-dimensional random feature regression. *arXiv preprint arXiv:2111.08234*, 2021.

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

A. Additional Details and Figures

A.1. Experimental Details for Digit Classification

In this section, we describe the details of the even vs odd handwritten digit classification experiment in Figure 2.

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

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

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.

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 2(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 three 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.

A.2. Not All Risks Have Monotonic Relationships

As discussed in the discussion around Theorem 2.1, the existence of a monotonic relation after distribution shifts is not a universal phenomenon that holds for broad classes of losses such as convex functions. We can demonstrate this via counterexample with the hinge and logistic losses that are used to train support vector machines, logistic regression models, and neural networks. We plot an example of this in Figure 3 for curves described by Lemma B.3 as a function of α only. In general, if both σ_E and α are free, monotonicity is even more difficult to achieve.

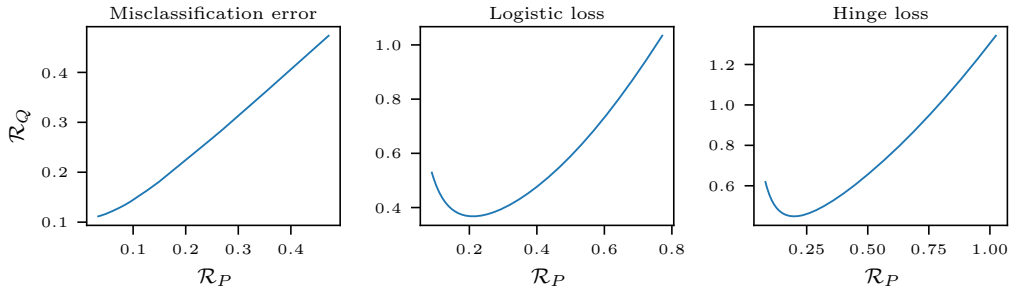

Figure 3: We compute, via Monte Carlo simulation with 10^5 random draws of (Z, W, E), the risk relationships for misclassification error (**left**) alongside the logistic loss $\psi(z^*, \hat{z}) = \log(1 + \exp(-\text{sign}(z^*)\hat{z}))$ (**middle**) and the hinge loss $\psi(z^*, \hat{z}) = \max\{1 - \text{sign}(z^*)\hat{z}\}$ (**right**). Here we have chosen $\rho = 5$, $\kappa_P = 0.1$, $\kappa_Q = 0.1$, $\kappa_R = 0.8$, $\sigma_\beta = 1$, $\varsigma = 1$, and fixed $\sigma_E = 1$. Unlike the misclassification error, these losses do not exhibit monotonic risk relationships as a function of α.

B. Proof of Theorem 2.1

In this section, we prove our main asymptotic result in Theorem 2.1. We first introduce a couple of definitions.

B.1. Definitions of Pseudo-Lipschitz Continuity and Convergence

We borrow the following definitions from Appendix A of Emami et al. (2020).

Definition B.1 (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 $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^r$,

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

Definition B.2 (Empirical convergence of a sequence)**.** Consider a sequence of collections of vectors $\mathcal{W}_n = (\mathbf{w}_i^{(n)})_{i=1}^n$, where $\mathbf{w}_i^{(n)} \in \mathbb{R}^r$. We say that the sequence $\mathcal{W}_n$ converges empirically with p-th order moments if there exists a random variable $W \in \mathbb{R}^r$ such that

(a) $\mathbb{E}\left[\|W\|_p^p\right] < \infty$ and

(b) for any $f\colon \mathbb{R}^r \to \mathbb{R}$ that is pseudo-Lipschitz continuous of order p,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}_i^{(n)}) = \mathbb{E}\left[f(W)\right].$$

With a slight abuse of notation, we will write

$$\mathbf{w}_i \overset{2}{\Rightarrow} W$$

to indicate empirical convergence of $\mathcal{W}_n$ to the random variable W for $p = 2$, omitting dependency of the convergence on n. This notion of convergence is equivalent to weak convergence plus convergence in p-th moment when elements of $\mathbf{w}_i$ are i.i.d. and is also equivalent to convergence in Wasserstein-p metric (Bayati & Montanari, 2011; Emami et al., 2020). For this reason, we also overload this notation, writing $X \overset{2}{\Rightarrow} Y$ for simultaneous weak convergence and convergence in second moment of a random vector $X \in \mathbb{R}^r$ to another random vector $Y \in \mathbb{R}^r$.

B.2. A Useful Lemma

Theorem 2.1 is a consequence of the following lemma, which shows that asymptotically, ground truth predictions and those of all solutions that use ridge regularization converge to relatively simple joint Gaussian distributions.

Lemma B.3. *For any estimate $\widehat{\boldsymbol{\beta}}$ solving (1), in the limit as $d \to \infty$, there exist $\alpha, \sigma_E > 0$ such that the test predictions $Z_P^* = \mathbf{x}^\top \boldsymbol{\beta}^*$ and $\widehat{Z}_P = \mathbf{x}^\top \widehat{\boldsymbol{\beta}}$ for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_P)$ and $Z_Q^*, \widehat{Z}_Q$ similarly defined for $\mathbf{x} = \tilde{\mathbf{x}} \circ \mathbf{s}, \tilde{\mathbf{x}} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_Q)$ converge almost surely as*

$$(Z_P^*, \widehat{Z}_P) \overset{2}{\Rightarrow} (Z, \alpha Z + E),$$
$$(Z_Q^*, \widehat{Z}_Q) \overset{2}{\Rightarrow} (\theta Z + W, \theta(\alpha Z + \varsigma E)),$$

where $Z \sim \mathcal{N}(0, (\kappa_P + \kappa_R)\sigma_\beta^2)$, $W \sim \mathcal{N}(0, \rho\kappa_Q\sigma_\beta^2)$, $E \sim \mathcal{N}(0, (\kappa_P + \kappa_R)\sigma_E^2)$, and $\theta = \sqrt{\rho\kappa_R/(\kappa_P + \kappa_R)}$.

To prove Lemma B.3, we leverage recent powerful results that have been proven recently for regularized ERM estimators in the proportional asymptotics regime. While we specifically use a result proven using the approximate message passing framework by Gerbelot et al. (2020) (see also a similar result by Emami et al. (2020)), we emphasize that the same result could be obtained via other techniques such as the convex Gaussian min-max theorem (Thrampoulidis et al., 2018)—we refer the reader to Salehi et al. (2019) for an example with logistic regression. All that is required is a result that shows that $\widehat{\boldsymbol{\beta}}$ converges to a linear combination of $\boldsymbol{\beta}^*$ and isotropic noise. Our strategy is to apply these results, which hold for rotationally invariant data distributions, to the restriction of the data space to the subspace corresponding to distribution P. Then using the characterization of the limiting joint distribution of $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$, we compute the joint distribution of model outputs.

B.2.1. Empirical Convergence of ERM Estimators

First, we re-state Theorem 1 of Gerbelot et al. (2020) in our notation. Strictly speaking, we state only the result of Lemma 4 of Gerbelot et al. (2020), since the full theorem requires an additional analytic continuation argument found in Appendix H of the same work. We refer the reader to that reference for details. We also remind the reader of the definition of the proximal operator for a function $f : \mathbb{R} \to \mathbb{R}$:

$$\mathrm{Prox}_f(x) = \arg\min_z f(z) + \frac{1}{2}(x - z)^2.$$

We now state the theorem, omitting assumptions that are naturally implied by the addition of the ridge penalty, and limiting the data assumption to isotropic Gaussian distributions.

Theorem B.4. *Consider the optimization problem*

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \ell(y_i, \mathbf{x}_i^\top \boldsymbol{\beta}) + \lambda' \sum_{j=1}^{d} \Omega(\beta_j) + \frac{\lambda^*}{2}\|\boldsymbol{\beta}\|_2^2, \tag{2}$$

where $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$, $y_i = \varphi(\mathbf{x}_i^\top \boldsymbol{\beta}^*, \xi_i)$, *and* $\xi_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. *Assume that*

- *the functions* ℓ *and* Ω *are proper, closed and convex functions;*
- *for any* $|z'| \leq C(1 + |z|)$ *for any* $z \in \mathrm{dom}(\ell(y, \cdot))$ *and* $z' \in \partial\ell(y, z)$ *for some universal* $C \geq 0$*, and the same holds for* Ω *on its domain;*
- *the labeling function* φ *is a proper, closed, continuous function;*
- *the empirical distribution of* $\boldsymbol{\beta}^*$ *converges empirically with second order moments, as defined in Definition B.2, to a zero-mean scalar random variable* B^* *with variance* σ_β^2*;*
- *the solution the set of fixed point equations (11) of Gerbelot et al. (2020) exists and is unique;*
- $n, d \to \infty$ *with fixed ratio* d/n.

Then there exist $a, c, \tilde{\sigma}_E > 0$ *and* $\lambda^* > 0$ *such that for any pseudo-Lipschitz function* ϕ *of order 2, the following holds almost surely*

$$\lim_{p\to\infty} \frac{1}{p}\sum_{j=1}^{p} \phi(\beta_j^*, \hat{\beta}_j) = \mathbb{E}\left[\phi\left(B^*, \mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)\right)\right],$$

where $\widetilde{\Omega}(b) = \lambda'\Omega(b) + \frac{\lambda^*}{2}b^2$ *and* $\widetilde{E} \sim \mathcal{N}(0, \tilde{\sigma}_E^2)$ *independent of* B^*.

The original result considers continuous labeling functions, but we can extend to bounded continuous functions with discontinuities of Lebesgue measure zero by upper and lower bounding by some continuous functions that get arbitrarily close to φ; we omit the details.

Specializing this result to the ridge penalty $\Omega(b) = \frac{1}{2}b^2$, for $\alpha = a/(1 + c(\lambda' + \lambda^*))$ and $\sigma_E = \tilde{\sigma}_E/(1 + c(\lambda' + \lambda^*))$, we have

$$(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}) \overset{2}{\Rightarrow} (B^*, \alpha B^* + E),$$

where $E \sim \mathcal{N}(0, \sigma_E^2)$ is independent of B^*.

B.2.2. EXTENDING TO THE SUBSPACE MODEL

Armed with the previous result, we can extend to the subspace model. First, we embed the isotropic model in Theorem B.4 into a space of dimension $d = d' + d_Q$ by appending d_Q dimensions such that $d_Q/d \to \kappa_Q$. Similarly, we partition the d' dimensions into d_P and d_R dimensions scaling with κ_P and κ_R, respectively, and set $c' = d'/d$. Define the projection operators $\mathbf{\Pi}_Q$ for the added dimensions and $\mathbf{\Pi}_P$ and $\mathbf{\Pi}_R$ for the original dimensions, respectively. Because the ridge regularization strength is positive, the learning problem in (1) with $\lambda = \lambda' + \lambda^*$ for $\mathbf{\Sigma}_P = \frac{1}{d}(\mathbf{\Pi}_P + \mathbf{\Pi}_R)$ has the same solution to that of (2) when restricted to the first d' coordinates, and the solution is forced to 0 for the last d_Q coordinates.

Now consider $\mathbf{x} \in \mathbb{R}^d$ divided into the corresponding sub-vectors $\mathbf{x}_P \in \mathbb{R}^{d_P}$, $\mathbf{x}_Q \in \mathbb{R}^{d_Q}$, and $\mathbf{x}_R \in \mathbb{R}^{d_R}$. Divide $\boldsymbol{\beta}^*$ into similarly partitioned $\boldsymbol{\beta}^*_P$, $\boldsymbol{\beta}^*_Q$, and $\boldsymbol{\beta}^*_R$, and $\widehat{\boldsymbol{\beta}}$ into $\widehat{\boldsymbol{\beta}}_P$ and $\widehat{\boldsymbol{\beta}}_R$ with $\widehat{\boldsymbol{\beta}}_Q = \mathbf{0}$. If $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$, then

$$(\mathbf{x}^\top \boldsymbol{\beta}^*, \mathbf{x}^\top \widehat{\boldsymbol{\beta}}) = (\mathbf{x}_P^\top \boldsymbol{\beta}^*_P + \mathbf{x}_Q^\top \boldsymbol{\beta}^*_Q + \mathbf{x}_R^\top \boldsymbol{\beta}^*_R, \mathbf{x}_P^\top \widehat{\boldsymbol{\beta}}_P + \mathbf{x}_R^\top \widehat{\boldsymbol{\beta}}_R).$$

Note that we have independence of these terms across P, Q, and R. Considering a particular subspace, observe that

$$\begin{aligned}(\mathbf{x}_P^\top \boldsymbol{\beta}^*_P, \mathbf{x}_P^\top \widehat{\boldsymbol{\beta}}_P) &\sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \frac{1}{d}\boldsymbol{\beta}^{*\top}_P \boldsymbol{\beta}^*_P & \frac{1}{d}\boldsymbol{\beta}^{*\top}_P \widehat{\boldsymbol{\beta}}_P \\ \frac{1}{d}\widehat{\boldsymbol{\beta}}^\top_P \boldsymbol{\beta}^*_P & \frac{1}{d}\widehat{\boldsymbol{\beta}}^\top_P \widehat{\boldsymbol{\beta}}_P \end{bmatrix}\right) \\ &\to \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \kappa_P \sigma_\beta^2 & \alpha\kappa_P\sigma_\beta^2 \\ \alpha\kappa_P\sigma_\beta^2 & \kappa_P(\alpha^2\sigma_\beta^2 + \sigma_E^2) \end{bmatrix}\right).\end{aligned}$$

Here convergence is in the sense that the covariance matrix converges almost surely thanks to pseudo-Lipschitz convergence of order 2 of $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$. This implies that

$$(\mathbf{x}_P^\top \boldsymbol{\beta}^*_P, \mathbf{x}_P^\top \widehat{\boldsymbol{\beta}}_P) \overset{2}{\Rightarrow} (Z_P, \alpha Z_P + E_P),$$

where $Z_P \sim \mathcal{N}(0, \kappa_P \sigma_P^2)$ and $E_P \sim \mathcal{N}(0, \kappa_P \sigma_E^2)$ are independent. An analogous result holds for $\mathbf{x}_R^\top \boldsymbol{\beta}^*_R$ and $\mathbf{x}_R^\top \widehat{\boldsymbol{\beta}}_R$ for $Z_R \sim \mathcal{N}(0, \kappa_R \sigma_\beta^2)$ and $E_R \sim \mathcal{N}(0, \kappa_R \sigma_E^2)$. Therefore, for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_P)$,

$$(\mathbf{x}^\top \boldsymbol{\beta}^*, \mathbf{x}^\top \widehat{\boldsymbol{\beta}}) \overset{2}{\Rightarrow} (Z_P + Z_R, \alpha(Z_P + Z_R) + E_P + E_R).$$

For distribution Q, observe that when $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_Q)$, $\mathbf{x}^\top \boldsymbol{\beta} = \widetilde{\mathbf{x}}^\top (\mathbf{s} \odot \boldsymbol{\beta})$ for $\widetilde{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \frac{\rho}{d}(\mathbf{\Pi}_Q + \mathbf{\Pi}_R))$. By assumption, $\mathbb{E}\left[S^2 B^{*2}\right] = \mathbb{E}\left[B^{*2}\right] = \sigma_\beta^2$, so there is no change to the effect of $\boldsymbol{\beta}^*$ for distribution Q. However, because $\mathbb{E}\left[S^2\right] = \varsigma^2$ which is not in general equal to one, there is a scaling of ς to the noise E. Adding the scaling factor ρ, we have for $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_Q)$ and $Z_Q \sim \mathcal{N}(0, \kappa_Q \sigma_\beta^2)$,

$$(\mathbf{x}^\top \boldsymbol{\beta}^*, \mathbf{x}^\top \widehat{\boldsymbol{\beta}}) \overset{2}{\Rightarrow} (\sqrt{\rho}(Z_R + Z_Q), \sqrt{\rho}(\alpha Z_R + \varsigma E_R)).$$

Letting $\theta = \sqrt{\rho\kappa_R/(\kappa_R + \kappa_P)}$, $Z = Z_P + Z_R$, $E = E_P + E_R$, and $W = \sqrt{\rho} Z_Q$, we obtain the stated result.

B.3. Proof of Theorem 2.1

The proof for each metric of Theorem 2.1 is different, so we present each case separately. Both are direct consequences of Lemma B.3.

B.3.1. MEAN SQUARED ERROR

For $\psi(z^*, \hat{z}) = (z^* - \hat{z})^2$, we have the following relationship:

$$\begin{aligned}\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) &= \mathbb{E}\left[((\theta Z + W) - \theta(\alpha Z + \varsigma E))^2\right] \\ &= \mathbb{E}\left[(\theta Z - \theta(\alpha Z + E + (\varsigma - 1)E))^2\right] + \mathbb{E}\left[W^2\right] \\ &= \theta^2 \mathbb{E}\left[(Z - (\alpha Z + E))^2\right] + \theta^2(\varsigma^2 - 1)\mathbb{E}\left[E\right] + \mathbb{E}\left[W^2\right] \\ &= \theta^2 \mathbb{E}\left[(\theta Z - \theta(\alpha Z + E))^2\right] + \theta^2(\varsigma^2 - 1)(\kappa_P + \kappa_R)\sigma_E^2 + \kappa_Q \sigma_\beta^2 \\ &= \theta^2 \mathcal{R}_P(\widehat{\boldsymbol{\beta}}) + \theta^2(\varsigma^2 - 1)(\kappa_P + \kappa_R)\sigma_E^2 + \kappa_Q \sigma_\beta^2.\end{aligned}$$

The dependence on σ_E^2 means that we do not have a relationship between $\mathcal{R}_Q$ and $\mathcal{R}_P$ in general. However, when the distribution shift is task-independent—that is, when $\varsigma = 1$—we have the simple linear relationship

$$\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) = \theta^2 \mathcal{R}_P(\widehat{\boldsymbol{\beta}}) + \kappa_Q \sigma_\beta^2.$$

B.3.2. Misclassification Error

For misclassification error, we first prove the following lemma.

Lemma B.5. *For two zero-mean jointly Gaussian random variables X and Y,*

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

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

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

The size of each of the intervals in the union is $\cos^{-1}\left(\mathbb{E}\left[\widetilde{X}\widetilde{Y}\right]\right)$, 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. □

For misclassification error, we can apply Lemma B.5:

$$\begin{aligned}
\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) &= \Pr((\theta Z + W)\theta(\alpha Z + \varsigma E) < 0) \\
&= \frac{1}{\pi}\cos^{-1}\left(\frac{\alpha\rho\kappa_R\sigma_\beta^2}{\sqrt{\left(\rho\kappa_R\sigma_\beta^2 + \rho\kappa_Q\sigma_\beta^2\right)\left(\alpha^2\kappa_R\sigma_\beta^2 + \varsigma^2\kappa_R\sigma_E^2\right)}}\right) \\
&= \frac{1}{\pi}\cos^{-1}\left(\frac{1}{\sqrt{\left(1 + \frac{\kappa_Q}{\kappa_R}\right)\left(1 + \frac{\varsigma^2\sigma_E^2}{\sigma_\beta^2\alpha^2}\right)}}\right).
\end{aligned}$$

By an analogous argument,

$$\mathcal{R}_P(\widehat{\boldsymbol{\beta}}) = \frac{1}{\pi}\cos^{-1}\left(\frac{1}{\sqrt{1 + \frac{\sigma_E^2}{\sigma_\beta^2\alpha^2}}}\right) = \frac{1}{\pi}\tan^{-1}\left(\frac{\sigma_E}{\sigma_\beta\alpha}\right).$$

Both of these functions are increasing functions of $\sigma_E/\sigma_\beta\alpha$, giving us the monotonically increasing relationship in the theorem statement. $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}})$ is also increasing in κ_Q/κ_R and ς. This means that if $\varsigma \geq 1$,

$$\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) > \mathcal{R}_P(\widehat{\boldsymbol{\beta}}).$$

Moreover, in the limit as $\kappa_Q/\kappa_R \to 0^+$, $\mathcal{R}_Q(\widehat{\boldsymbol{\beta}}) = \frac{1}{\pi}\tan^{-1}\left(\frac{\varsigma\sigma_E}{\sigma_\beta\alpha}\right)$, giving us the final linear relationship.

B.4. Extension to General Separable Regularization Penalties

By Theorem B.4, the true parameter $\boldsymbol{\beta}^*$ and its estimate $\widehat{\boldsymbol{\beta}}$ empirically converge as

$$(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}) \overset{2}{\Rightarrow} \left(B^*, \mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)\right)$$

for some $a, c > 0$ and $\widetilde{E} \sim \mathcal{N}(0, \tilde{\sigma}_E^2)$ that is independent of B^*. However, observe that in order to prove Lemma B.3, we only need to determine the limits of terms of the form $\frac{1}{d'}\widehat{\boldsymbol{\beta}}_P^\top\boldsymbol{\beta}_P^*$ and $\frac{1}{d'}\boldsymbol{\beta}_P^{*\top}\boldsymbol{\beta}_P^*$. As $d \to \infty$, we can find $\alpha, \sigma_E > 0$ such that:

$$\frac{1}{d'}\widehat{\boldsymbol{\beta}}_P^\top\boldsymbol{\beta}_P^* \to \kappa_P\,\mathbb{E}\left[B^*\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)\right] = \kappa_P\alpha\sigma_\beta^2$$
$$\frac{1}{d'}\boldsymbol{\beta}_P^{*\top}\boldsymbol{\beta}_P^* \to \kappa_P\,\mathbb{E}\left[\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)^2\right] = \kappa_P(\alpha^2\sigma_\beta^2 + \sigma_E^2).$$

Concretely, we can use

$$\alpha = \frac{\mathbb{E}\left[B^*\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)\right]}{\sigma_\beta^2}, \quad \sigma_E^2 = \mathbb{E}\left[\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)^2\right] - \alpha^2\sigma_\beta^2.$$

As long as $s(b) = 1$, there is no further change, and the rest of the proof follows. Following Gerbelot et al. (2020), we can let $\lambda^* \to 0$ using an analytic continuation when sufficient regularity assumptions on ℓ, Ω, and ψ are met in order to remove the ridge regression component, extending our result to general separable penalties. However, for general s, the problems are not equivalent to our previous formulation; instead of simply having a fixed $\mathbb{E}\left[s(B^*)^2\right]$, we must also evaluate

$$\mathbb{E}\left[s(B^*)^2B^*\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)\right] \quad \text{and} \quad \mathbb{E}\left[s(B^*)^2\mathrm{Prox}_{c\widetilde{\Omega}}\left(aB^* + \widetilde{E}\right)^2\right],$$

which will not simplify to functions of α, σ_E^2, and ς in general.