Project with Source, Probe with Target: Extracting Useful Features for Adaptation to Distribution Shifts

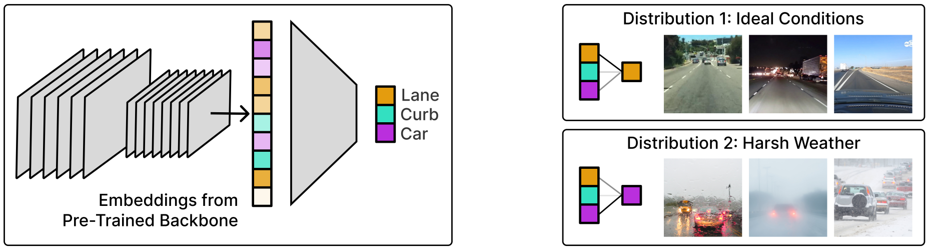
Annie S. Chen*
Stanford University

Yoonho Lee*
Stanford University

Amrith Setlur
Carnegie Mellon University

Sergey Levine
UC Berkeley

Chelsea Finn
Stanford University


Embeddings from
Pre-Trained Backbone
Lane
Curb
Car
Distribution 1: Ideal Conditions
Distribution 2: Harsh Weather

(a) **Pro**ject with Large Source Dataset

(b) **Pro**be with Small Target Dataset

Figure 1: **The Project and Probe (Pro^2) framework for adapting to different target distributions.** (a) We first use a large source dataset to project pre-trained feature embeddings onto a set of predictive features while enforcing orthogonality. (b) For a new target distribution, we learn a linear layer on top of the projected features. This step adaptively chooses features in a data-efficient manner.

1 Introduction

Datasets often exhibit a spurious correlation, where a *shortcut feature* that is predictive on the training data can be misleading on a shifted distribution of inputs, because it does not capture the underlying causal relationships. However, identifying the true causal features can be notably difficult or simply not possible in some scenarios. Furthermore, we argue that approximating only the causal features may not always be the best approach: *shortcut features can be useful in some situations*. Prior work has studied this in human decision making: non-causal mental shortcuts and heuristics can sometimes be more effective than making a logical deduction from all available information (Tversky and Kahneman, 1974; Simon et al., 1989; Gigerenzer and Gaissmaier, 2011). As an example, consider an autonomous vehicle tasked with following a lane. While the ground-truth causal feature for lane following is the road markings, the position of other cars in the lane, a "shortcut feature", is also predictive of the lane following task. In conditions where the causal feature is less informative (e.g., road markings not visible due to fog), it can be best to rely on other features (e.g., follow the car in front). Therefore, in this work, we aim to extract a variety of potentially useful features and identify which ones to use for a given situation.

Recent works have found that failures due to spurious correlations can be addressed at test time by retraining a final linear head (Rosenfeld et al., 2022; Kirichenko et al., 2022; Mehta et al., 2022). These methods demonstrate that such adaptation reliably improves performance with even a small amount of additional target data. However, it is not clear whether linear probing is the most sample-efficient way to adapt to new distributions, as the features may contain redundant or non-predictive/noisy information. We highlight an important but underexplored insight for these adaptation methods: the learned head should be able to extract the most suitable features for varied target distributions, which may include both shortcut and robust features, and choose between them to best adapt to a particular target distribution.

We propose Project and Probe (Pro^2), a simple, computationally efficient, and data-efficient method for adapting to unknown target distributions. Pro^2 first learns a projection of pre-trained embedding vectors, which is optimized to extract a diverse set of features that are each predictive of labels. More specifically, we first use a source dataset to project pre-trained feature embeddings onto a set of predictive features while enforcing orthogonality to ensure that each projected dimension holds information not present in other dimensions. We expect this learned feature space to compactly contain a diverse set of predictive features while discarding non-predictive or redundant information. PRO^2 then learns a linear head to interpolate between the projected features. Both the linear projection and head require minimal computational overhead, making PRO^2 a practical method for adapting to new target distributions. Figure 1 shows a visual summary of PRO^2.

To support our approach, we provide a theoretical analysis, which shows how the projection matrix learned by PRO^2 is optimal in an information-theoretic sense, resulting in better generalization in the low-data regime due to a favorable bias-variance tradeoff. We conduct experiments on a variety of distribution shift settings across 4 datasets. We find that standard linear probing is relatively inefficient as an approach to adapting to target data, while PRO^2 substantially improves sample efficiency in this setting. Our results show that given limited target data, PRO^2 is consistently competitive with standard debiasing methods that attempt to directly learn a robust classifier, while outperforming them by 5-15% on data distributions in which the shortcut feature is more useful.

2 RELATED WORK

Adapting to distribution shifts. Recent works have proposed various methods for adapting models at test time with some labeled target data Sun et al. (2020); Varsavsky et al. (2020); Iwasawa and Matsuo (2021); Wang et al. (2020); Zhang et al. (2021); Gandelsman et al. (2022); Lee et al. (2022a). In particular, given a feature embedding produced by a pretrained network with sufficient expressivity, training a final linear head, also known as linear probing, suffices for adapting to datasets with spurious correlations Kirichenko et al. (2022); Mehta et al. (2022); Izmailov et al. (2022) as well as in the setting of domain generalization Rosenfeld et al. (2022). As detailed further in Section 3, we specifically focus on scenarios in which we have very little target data (only $4 \sim 256$ datapoints). We find that in this setting, training a final linear head in the default manner is not the most data-efficient way to adapt. PRO^2, which breaks this training down into 2 steps, is able to more effectively extract useful features and interpolate between them for varying target distributions, leading to improved sample efficiency with limited target data.

Learning diverse features for spurious datasets. Neural networks tend to be biased towards learning simple functions that rely on shortcut features (Arpit et al., 2017; Gunasekar et al., 2018; Shah et al., 2020; Geirhos et al., 2020; Pezeshki et al., 2021; Li et al., 2022; Lubana et al., 2022). To better handle novel distributions, it is important to consider the entire set of functions that are predictive on the training data (Fisher et al., 2019; Semenova et al., 2019; Xu et al., 2022). Recent diversification methods discover such a set (Teney et al., 2022; Lee et al., 2022b; Pagliardini et al., 2022). The latter two methods use additional assumptions such as unlabeled data and we find that PRO^2 outperforms the former in Section 6.

Compression & feature selection. In aiming to extract important features and discarding repetitive information, PRO^2 is related to work on compression May et al. (2019) and information bottlenecks Tishby et al. (2000); Alemi et al. (2016). Our method is also closely related to methods that learn projections such as principal component analysis (PCA) and linear discriminant analysis (LDA). Beyond these representative methods, there is an immense body of work on feature selection (Dash and Liu, 1997; Liu and Motoda, 2007; Chandrashekar and Sahin, 2014; Li et al., 2017) and dimensionality reduction (Lee et al., 2007; Sorzano et al., 2014; Cunningham and Ghahramani, 2015). Among all projection-based methods, LDA is the most related to ours, but it only learns the single most discriminative direction. In Corollary 8, we show that PRO^2 with dimensionality $d = 1$ provably recovers the LDA direction in a shifted homoscedastic Gaussian model, and that using higher values of d is critical in adapting to higher degrees of distribution shift. Generally, most methods (including LDA) operate in the setting without distribution shift.

3 ADAPTATION TO DISTRIBUTION SHIFT

We now describe our problem setting, where the goal is to adapt a model so as to provide an accurate decision boundary under distribution shift given a limited amount of target distribution information.

We consider a source distribution $p_S(x, y)$ and multiple target distributions $p_T^1(x, y), p_T^2(x, y), \cdots$. The source dataset $\mathcal{D}_S \in (\mathcal{X} \times \mathcal{Y})^N$ is sampled from the source distribution p_S. We evaluate adaptation to each target distribution p_T^i given a small set of labeled target data $\mathcal{D}_T^i \in (\mathcal{X} \times \mathcal{Y})^M$, where $M \ll N$ so the model must learn from both the source and target data for best performance. We measure the post-adaptation average accuracy of the model on a held-out target dataset from the same distribution p_T^i. We note that our setting differs from two settings studied in prior works; we discuss these differences in Appendix B.

4 PROJECT AND PROBE

We now describe PRO^2, a framework for few-shot adaptation to distribution shifts. PRO^2 is composed of two steps: (1) learn a projection Π that maps pre-trained embeddings onto orthogonal directions, and (2) learn a classifier g using projected embeddings. Before Step (1), we use a pre-trained backbone model $f : \mathcal{X} \to \mathbb{R}^D$ to map the datapoints to D-dimensional embeddings. This backbone model extracts meaningful features from the raw inputs, resulting in a low-dimensional embedding space, for example $224 \times 224 \times 3$ images to $D = 1024$-dimensional embeddings.

Algorithm 1 Project and Probe

```
  Input: Source data $\mathcal{D}_S$, Target data $\mathcal{D}_T$,
  Backbone $f : \mathcal{X} \to^D$
  Initialize $\Pi :^D \to^d$
  for $i$ in $1 \dots d$ do
      $\Pi_i \leftarrow \arg\min \mathcal{L}_S(\Pi_i(f(x)), y)$
          subject to $\Pi_j \perp \Pi_i$ for all $j < i$
  Initialize $g :^d \to \mathcal{Y}$
  $g \leftarrow \arg\min \mathcal{L}_T(g(\Pi(f(x))), y)$
```

Step 1: Project with source. Recall that we operate in the few-shot setting, where we may have fewer target datapoints than even embedding dimensions ($M < D$). We would like to select a suitable decision boundary by *interpolating* over a basis of decision boundaries, which is mathematically identical to selecting a set of linear features. Thus, the question we must answer is: which set of linear features of the D-dimensional feature space should we retain? First, it should be clear that the features should form an orthogonal basis, as otherwise they will be redundant. Second, the features should be discriminative, in the sense that they are sufficient to solve the desired prediction task. Lastly, there should not be too many of them, since the more features we include (i.e., the larger the rank of the basis we learn), the more samples we'll need from the target domain to find the best decision boundary in the corresponding set.

To learn a feature space that satisfies these desiderata, we parameterize a linear projection $\Pi : \mathbb{R}^D \to \mathbb{R}^d$ that maps the embeddings to a reduced space ($d \leq D$). Specifically, we use the source data to learn a complete orthonormal basis for the embedding space $\Pi_1, \Pi_2, \dots, \Pi_d \in^D$, by learning each basis vector with the constraint that it is orthogonal to all vectors before it:

$$\Pi_i = \arg\min \mathbb{E}_{(x,y)\sim\mathcal{D}_S} \mathcal{L}(\Pi_i(f(x)), y) \quad \text{s.t.} \quad \Pi_j \perp \Pi_i \text{ for all } j < i. \tag{1}$$

Note that this induces a natural ranking among the basis vectors. This collection of orthogonal vectors constitute the rows of our projection matrix Π. In our implementation, we do projected gradient descent, enforcing orthogonality using QR decomposition on the projection matrix after every gradient step. See Appendix D for pseudocode on this step. We find that it is particularly beneficial to use a small $d \ll D$, even $d = 1$, in when adapting to small distribution shifts and use larger d for more severe distribution shifts.

Step 2: Probe with target. After learning Π, we learn a classifier $g : \mathbb{R}^d \to \mathcal{Y}$ that maps projected embeddings to target labels: $g = \arg\min_g \mathbb{E}[\mathcal{L}((g \circ \Pi \circ f)(x), y)]$. Since the projection Π was optimized to a diverse set of the most discriminative features for the source data, we expect the initial projected features to be particularly predictive when the distribution shift is relatively small.

In summary, PRO^2 is a simple and lightweight framework that addresses the problem of few-shot adaptation in the presence of distribution shifts; we summarize its overall structure in Algorithm 1. In our implementation, we use cached embeddings for all source and target datapoints, such that feeding raw inputs through f is a one-time cost that is amortized over epochs and experiments, making our framework scalable and efficient. As an orthogonal improvement to our work, one could additionally fine-tune the backbone network on source data. In Section 5, we theoretically analyze the properties of the projection and classifier learned by PRO^2. We then empirically evaluate PRO^2 on a variety of distribution shifts in Section 6.

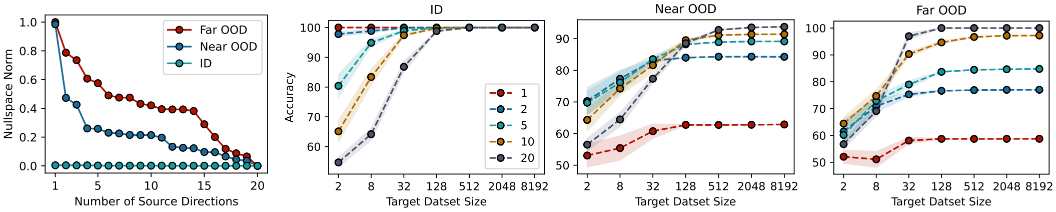


Figure 2: **Evaluation of $\text{P}\textsc{ro}^2$ on shifted homoscedastic Gaussian data.** (Left) The x- and y-axes denote dimensionality of A_d and nullspace norm, respectively. Nullspace norm drops slowly for more severe distribution shifts. (Right) For less severe distribution shifts (ID and Near OOD), low-dimensional projections suffer from less bias, resulting in higher accuracy in the low-data regime. For the Far OOD distribution, using all 20-dimensional features is best, as bias drops more slowly.

5 Analysis

In this section, we present a theoretical analysis of $\text{P}\textsc{ro}^2$, aiming to understand how our proposed orthogonal feature selection procedure can lead to sample-efficient adaptation under distribution shifts. Intuitively, the more shift we can expect, the more features we should need to adapt to it, which in turn requires more samples during adaptation (to fit the features accurately). However, the choice of how we extract features influences the rate at which the sample complexity grows under distribution shift: while large shifts may still require many features, if the features are prioritized well, then smaller shifts might require only a very small number of features, and thus require fewer samples.

In our analysis, we first show that $\text{P}\textsc{ro}^2$ learns a projection matrix that is optimal in an information-theoretic sense. We next show that using fewer features (d) leads to lower variance, which scales as $(\mathcal{O}(\sqrt{d/M}))$ given M target samples, but at a cost in bias, whichin some cases scales as $\mathcal{O}(\sqrt{1-(d/D)} \cdot \mathrm{KL}(p_S||p_T))$ and grows with the amount of distributional shift between p_S and p_T. In Section 5.1, we first analyze the specific features learned by $\text{P}\textsc{ro}^2$ with minimal distributional assumptions. Then, in Section 5.2, we specialize these more general results to a shifted homoscedastic Gaussian (SHOG) model, and demonstrate the bias-variance tradeoff empirically. Additional theoretical results and proofs for the results in this section can be found in Appendix C.

5.1 Bias-variance tradeoffs for general shifts.

From the original D-dimensional feature representations given by our feature backbone f, we want our learned linear projections $\Pi :^{D} \rightarrow^{d}$ to retain as much information as possible that is relevant in predicting the label y. In other words, we want to maximize the mutual information between the projected features $\Pi(\mathbf{x})$ and the labels y. We first formally characterize the solution found by the projection step in $\text{P}\textsc{ro}^2$ as maximizing this mutual information amongst all rank d matrices with orthogonal columns.

Theorem 1 (Information in projected input)**.** *When the distributions $p((\mathbf{x} - \mathbb{E}[\mathbf{x}]) \mid \mathrm{y})$ are identical for each y. the solution $\{\Pi_i\}_{i=1}^{d}$ returned by $\text{P}\textsc{ro}^2$ maximizes the mutual information $I(\mathbf{A}\mathbf{x};\ \mathrm{y})$ (and a strict upper bound on it otherwise) among all $D \times d$ matrices $\mathbf{A}$ with orthogonal columns.*

This theorem shows that the projection matrix Π learned by $\text{P}\textsc{ro}^2$ is optimal in an information-theoretic sense of retaining the most information about y, on the source distribution. This is in line with the motivation for the orthogonality constraint, which was to minimize redundancy while gathering different features that are each predictive of the label on source. Next, we analyze the properties of Π on the *target distribution* to understand how the degree of distributional shift affects sample efficiency during adaptation.

Probing on the target distribution. We first introduce some additional notation specific to the target distribution. For some projection Π, let $\mathbf{\Pi}_d$ denote the projection matrix for $\mathrm{span}(\{\Pi_i\}_{i=1}^{d})$, $\mathbf{\Pi}_d{=}[\Pi_1, .., \Pi_d][\Pi_1, .., \Pi_d]^\top$. Denote the target error for classifier $\mathbf{w}$ as $\mathcal{L}_T(\mathbf{w}) \triangleq \mathbb{E}_{p_T} l(\langle \mathbf{w}, \mathbf{x} \rangle,\ \mathrm{y})$, and the bias incurred by probing over the projected features $\mathrm{span}(\{\Pi_i\}_{i=1}^{d})$ as: $b_d \triangleq \min_{\mathbf{w}' \in \mathrm{span}(\{\Pi_i\}_{i=1}^{d})} \mathcal{L}_T(\mathbf{w}') - \min_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_T(\mathbf{w})$. We also denote the d-

dimensional weight vector learned by PRO^2 on the M projected target samples as: $\hat{\mathbf{w}}_d \triangleq \min_{\substack{\mathbf{w}\in\text{span}(\{\Pi_i\}_{i=1}^d) \\ \|\mathbf{w}\|_2\leq 1}} \sum_{i=1}^M l(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle,\ \mathrm{y}^{(i)})$.

In Theorem 2, we describe the full bias-variance tradeoff where we see that the variance term is also controlled by the number of features d but unlike the bias is independent of the nature of shift between source and the target.

Theorem 2 (bias-variance tradeoff)**.** *When the conditions in Lemma 5 hold and when $\|\mathbf{x}\|_\infty = \mathcal{O}(1)$, for B-bounded loss l, w.h.p. $1-\delta$, the excess risk for the solution $\hat{\mathbf{w}}_d$ of* PRO^2 *that uses d features is $\mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w})$*

$$\|(\mathbf{I}_D - \mathbf{\Pi}_d)\mathbf{w}_T^*\|_2 + \frac{\sqrt{d} + B\sqrt{\log(1/\delta)}}{\sqrt{M}}, \tag{2}$$

where the first term controls the bias and the second controls the variance.

This result provides insights on what factors affect generalization when probing on target data. Tighter compression of the original representation, i.e., using a smaller d, increases bias while decreasing variance. The rate of bias increase is determined by the degree of distribution shift, where more severe shifts correspond to a steeper increase in bias. The distribution shift has no effect on variance, and variance can only be decreased by using a low-dimensional represent (at the cost of bias) or learning from a larger dataset.

5.2 Bias-variance tradeoff in shifted Gaussian model.

In this subsection, we consider a simplified setting of a shifted homoscedastic Gaussian (SHOG). Within this model, we show that the more general statement in Theorem 2 can be simplified further to provide a more intuitive relationship between the factors that affect generalization. Furthermore, we empirically demonstrate the behavior predicted by our bounds on synthetic SHOG data.

Shifted homoscedastic Gaussian (SHOG) model of distribution shift. We model the source distribution as a Bernoulli mixture model of data in which binary labels are balanced ($\mathrm{y} \sim \text{Bern}(0.5)$) and the class conditional distributions are homoscedastic multi-variate Gaussians:

$$\mathbf{x} \mid \mathrm{y} \sim \mathcal{N}(\mu_{\mathrm{y}}, \Sigma_S) \quad \text{for} \quad \mathrm{y} \in \{0, 1\},$$

where $\mu_1, \mu_2 \in^D$ are mean vectors and $\Sigma_S \in^{D\times D}$ is the shared covariance matrix. The target distribution has the same label distribution and Gaussian means, but a different covariance matrix given by Σ_T. We study how the relation between the two covariance matrices Σ_S, Σ_T can affect the bias term b_d when Π_d is either returned by PRO^2 or a random projection matrix with columns drawn uniformly over the sphere S^{d-1}.

We specialize the more general bias-variance tradeoff result to a shifted homoscedastic Gaussian (SHOG) model in Corollary 3, where we derive a simpler bound characterizing the tradeoff between performance, the value of d, and the amount of distributional shift.

Corollary 3 (tradeoff under SHOG)**.** *Under our SHOG model of shift, and conditions for a random projection $\mathbf{\Pi}_d$ in Lemma 9, the target error $\mathcal{L}_T(\hat{\mathbf{w}}_d)\mathcal{O}\sqrt{1-\frac{d}{D}} \cdot \text{KL}(p_S||p_T) + \sqrt{\frac{d}{M}}$, when $\Sigma_T = O(1)$.*

In Figure 2, we plot the nullspace norm $\|\Sigma_S\|_{\text{op}}$ for different d in three target distributions of varying distribution shift severity in the SHOG model. We see that the more severe shifts have a higher norm, indicating that the OOD distributions suffer from high bias when d is low. Indeed, we see that the ID distribution suffers from virtually no bias, making $d = 1$ achieve highest target accuracy for all dataset sizes. In contrast, the Near OOD and Far OOD distributions suffer from high bias of up to 40% accuracy, and higher projection dimension d is needed for adaptation, as predicted by Corollary 3.

6 Experiments

In this section, we aim to empirically answer the following questions: (1) Can PRO^2 identify a feature-space basis for rapid adaptation, and how does it compare to other methods for extracting

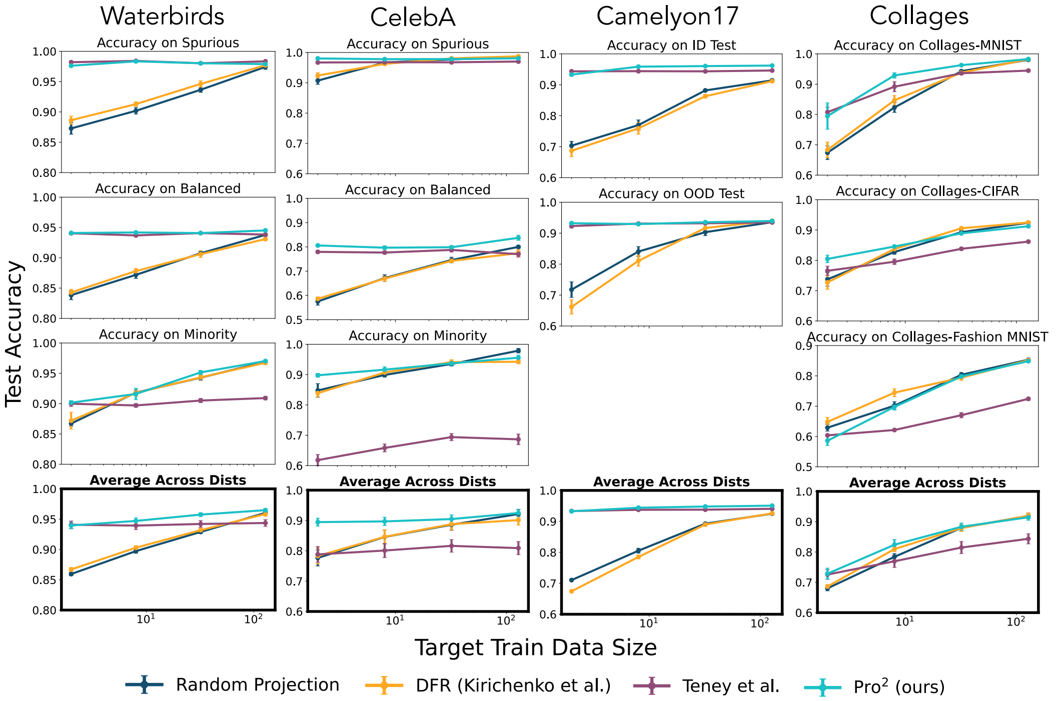


Figure 3: **Main results.** We compare 4 different methods for learning features to adapt to a target distribution: (1) Random Projection, (2) DFR Kirichenko et al. (2022), (3) Teney et al. (2021), and (4) PRO^2. We report the mean and standard error of target accuracy across 10 random seeds. PRO^2 is the best performing or tied for best performing method *across all datasets and dataset size*, while also substantially outperforming Random Projection and DFR in the low-data regime on all settings. PRO^2 also outperforms Teney et al. (2021) on average on 3 of the 4 datasets particularly when given

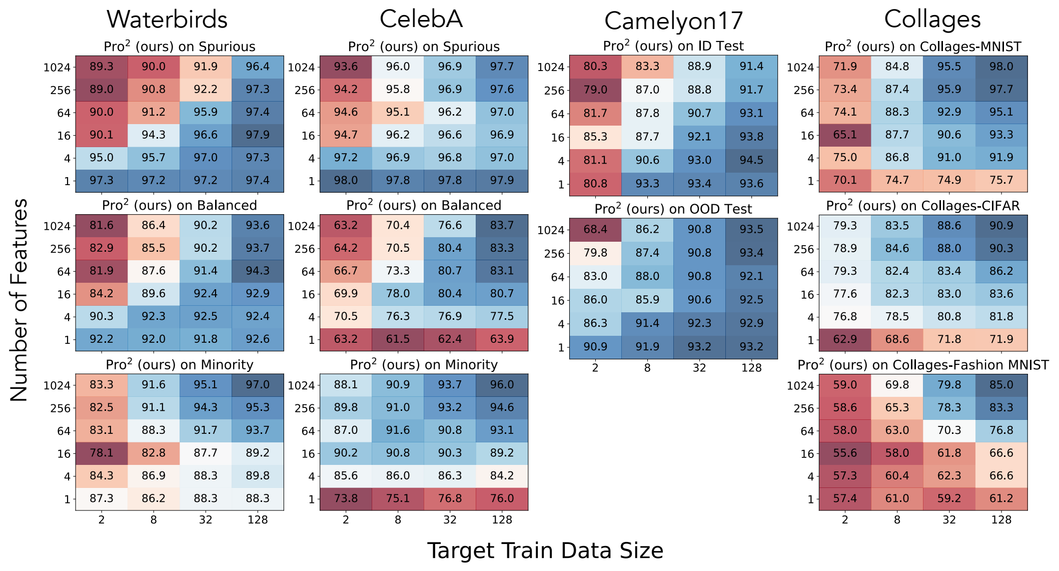


Figure 4: **Projection dimensionality of PRO^2 and severity of distribution shift.** We vary the dimensions d (y-axis) of PRO^2 and report target accuracy after training on target datasets of different size (x-axis) on our 4 datasets. Higher d was required to adapt to more severe shifts, while for milder shifts, lower d sometimes results in higher accuracy, as can be seen in the Spurious distribution of Waterbirds/CelebA.

features? (2) How does the dimensionality of the feature-space basis affect sample efficiency in different distribution shift conditions? We provide additional empirical results and analyses, such as showing that the adaptation performance of PRO^2 improve with better pre-trained backbones, in Appendix E. Details on pre-trained models and training details are in Appendix D.

6.1 Experimental Setup

Datasets. We run experiments on four datasets with distribution shifts: 4-way Collages (Teney et al., 2021), Waterbirds (Sagawa et al., 2020), CelebA (Liu et al., 2015), and Camelyon (Bandi et al., 2018) datasets. Each of these datasets have a source distribution that we use for training and multiple target distributions for evaluation. For all settings, we use the original source datasets, which each contain thousands of datapoints. For target data, we subsample very small label-balanced datasets for adaptation, with $\{2, 8, 32, 128\}$ images per label. The remaining target distribution datapoints are used for evaluation. Due to space constraints, we describe the different target distributions in Appendix D.

Computational efficiency. Similarly to Mehta et al. (2022), we use feature embeddings from a pre-trained backbone without fine-tuning. Our aim is to develop methods that can leverage pretrained models out-of-the-box with minimal computational requirements: our training involves at most two linear layers on top of cached feature vectors. For all comparisons, we hyperparameter tune over three learning rates (0.1, 0.01, and 0.001) and three L_2 regularization weights (0.1, 0.01, 0.001). In our main experiments in Section 6.2, we also sweep over six projection dimensions ($d = 1, 4, 16, 64, 256, 1024$) and report results over 10 runs. As a demonstration of the computational efficiency of Pro^2, after caching pre-trained embeddings, we collectively ran all experiments in Section 6.2, which is nearly $30,000$ runs due to hyperparameter tuning, within 24 *hours* using four standard CPUs and *no GPUs*.

6.2 Comparison to prior projection methods

We investigate whether Pro^2 can extract features that can facilitate adaptation to different distribution shifts, and how it compares other feature extraction methods. We perform a comprehensive experimental evaluation on the four datasets, comparing Pro^2 against three other projection methods: (1) Random Projection, (2) DFR Kirichenko et al. (2022), i.e., linear probing, and (3) Teney et al. (2021), which learns diverse features by minimizing gradient similarity. Experiments in Figure 3 indicate that across all distributions and datasets, Pro^2 significantly outperforms Random Projection and DFR, especially in the low-data regime. In particular, these results show that linear probing, the strategy adopted by several additional prior works by default Mehta et al. (2022); Izmailov et al. (2022), is a suboptimal strategy for few-shot adaptation, likely because raw embeddings contain redundant or non-informative information. Teney et al. (2021) is sufficient in some scenarios with milder distribution shift, but fails given large datasets or severe distribution shifts. In contrast, Pro^2 improves sample efficiency while remaining competitive across all settings. This indicates that the feature diversity from the orthogonality constraint gives Pro^2 better coverage of different features, enabling better adaptation to severe distribution shifts given enough target data.

6.3 Projection dimension and shift severity

We investigate how the feature-space dimension d in Pro^2 affects sample efficiency for different degrees of distribution shift. Experiments in Figure 4 show that when the distribution shift is less severe, such as the Spurious distributions on Waterbirds and CelebA, it is helpful to reduce the number of features used. Fewer features suffice here because the top-ranked features from the source data are also predictive on the target distribution. However, when the distribution shift is more severe, such as the Minority distributions on Waterbirds and CelebA or Collages-Fashion MNIST and Collages-CIFAR, it is helpful to increase the number of features used. These empirical results are supported formally by our theoretical results in Section 5, which show that the optimal number of features to use increases with distribution shift severity.

References

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2016). Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arpit, D., Jastrzębski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y., et al. (2017). A closer look at memorization in deep networks. In *International Conference on Machine Learning*.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. (2018). From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Cover, T. M. and Thomas, J. A. (1991). Information theory and statistics. *Elements of information theory*, 1(1):279–335.

Creager, E., Jacobsen, J.-H., and Zemel, R. (2021). Environment inference for invariant learning. In *International Conference on Machine Learning*.

Cunningham, J. P. and Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *The Journal of Machine Learning Research*, 16(1):2859–2900.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(1-4):131–156.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Gandelsman, Y., Sun, Y., Chen, X., and Efros, A. A. (2022). Test-time training with masked autoencoders. *arXiv preprint arXiv:2209.07522*.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Gigerenzer, G. and Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62(1):451–482.

Gunasekar, S., Lee, J. D., Soudry, D., and Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Iwasawa, Y. and Matsuo, Y. (2021). Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. (2022). On feature learning in the presence of spurious correlations. *arXiv preprint arXiv:2210.11369*.

Kakade, S. M., Sridharan, K., and Tewari, A. (2008). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21.

Kirichenko, P., Izmailov, P., and Wilson, A. G. (2022). Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

Koh, P. W., Sagawa, S., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., Lee, T., et al. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

Kornblith, S., Shlens, J., and Le, Q. (2018). Do better imagenet models transfer better? arxiv 2018. *arXiv preprint arXiv:1805.08974*.

Kumar, A., Raghunathan, A., Jones, R., Ma, T., and Liang, P. (2022). Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*.

Lee, J. A., Verleysen, M., et al. (2007). *Nonlinear dimensionality reduction*, volume 1. Springer.

Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. (2022a). Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466*.

Lee, Y., Yao, H., and Finn, C. (2022b). Diversify and disambiguate: Learning from underspecified data. *arXiv preprint arXiv:2202.03418*.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., and Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45.

Li, Z., Evtimov, I., Gordo, A., Hazirbas, C., Hassner, T., Ferrer, C. C., Xu, C., and Ibrahim, M. (2022). A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Liu, H. and Motoda, H. (2007). *Computational methods of feature selection*. CRC press.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lubana, E. S., Bigelow, E. J., Dick, R. P., Krueger, D., and Tanaka, H. (2022). Mechanistic mode connectivity. *arXiv preprint arXiv:2211.08422*.

May, A., Zhang, J., Dao, T., and Ré, C. (2019). On the downstream performance of compressed word embeddings. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Mehta, R., Albiero, V., Chen, L., Evtimov, I., Glaser, T., Li, Z., and Hassner, T. (2022). You only need a good embeddings extractor to fix spurious correlations.

Morwani, D., Batra, J., Jain, P., and Netrapalli, P. (2023). Simplicity bias in 1-hidden layer neural networks. *arXiv preprint arXiv:2302.00457*.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. (2020). Learning from failure: Training debiased classifier from biased classifier. *Conference on Neural Information Processing Systems*.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Pagliardini, M., Jaggi, M., Fleuret, F., and Karimireddy, S. P. (2022). Agree to disagree: Diversity through disagreement for better transferability. *arXiv preprint arXiv:2202.04414*.

Petridis, S. and Perantonis, S. J. (2004). On the relation between discriminant analysis and mutual information for supervised linear feature extraction. *Pattern Recognition*, 37(5):857–874.

Pezeshki, M., Kaba, S.-O., Bengio, Y., Courville, A., Precup, D., and Lajoie, G. (2021). Gradient starvation: A learning proclivity in neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

Rosenfeld, E., Ravikumar, P., and Risteski, A. (2022). Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*.

Semenova, L., Rudin, C., and Parr, R. (2019). A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P., and Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Conference on Neural Information Processing Systems*.

Sharif Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

Simon, H. A. et al. (1989). The scientist as problem solver. *Complex information processing: The impact of Herbert A. Simon*, pages 375–398.

Sorzano, C. O. S., Vargas, J., and Montano, A. P. (2014). A survey of dimensionality reduction techniques. *arXiv preprint arXiv:1403.2877*.

Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., and Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR.

Teney, D., Abbasnejad, E., Lucey, S., and Hengel, A. v. d. (2021). Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. *arXiv preprint arXiv:2105.05612*.

Teney, D., Abbasnejad, E., Lucey, S., and van den Hengel, A. (2022). Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16761–16772.

Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

Varsavsky, T., Orbes-Arteaga, M., Sudre, C. H., Graham, M. S., Nachev, P., and Cardoso, M. J. (2020). Test-time unsupervised domain adaptation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 428–436. Springer.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. (2020). Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wortsman, M., Ilharco, G., Kim, J. W., Li, M., Kornblith, S., Roelofs, R., Lopes, R. G., Hajishirzi, H., Farhadi, A., Namkoong, H., and Schmidt, L. (2022). Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7959–7971.

Xu, Y., He, H., Shen, T., and Jaakkola, T. (2022). Controlling directions orthogonal to a classifier. *arXiv preprint arXiv:2201.11259*.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A. S., Neumann, M., Dosovitskiy, A., et al. (2019). A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. (2021). Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678.

Zhang, M. and Ré, C. (2022). Contrastive adapters for foundation model group robustness. *arXiv preprint arXiv:2207.07180*.

A Abstract

Conventional approaches to robustness try to learn a model based on causal features. However, identifying maximally robust or causal features may be difficult in some scenarios, and in others, non-causal "shortcut" features may actually be more predictive. We propose a lightweight, sample-efficient approach that learns a diverse set of features and adapts to a target distribution by interpolating these features with a small target dataset. Our approach, Project and Probe (Pro^2), first learns a linear projection that maps a pre-trained embedding onto orthogonal directions while being predictive of labels in the source dataset. The goal of this step is to learn a variety of predictive features, so that at least some of them remain useful after distribution shift. Pro^2 then learns a linear classifier on top of these projected features using a small target dataset. We theoretically show that Pro^2 learns a projection matrix that is optimal for classification in an information-theoretic sense, resulting in better generalization due to a favorable bias-variance tradeoff. Our experiments on eight distribution shift settings show that Pro^2 improves performance by 5-15% when given limited target data compared to prior methods such as standard linear probing.

B Comparison to Problem Setting in Existing Works

Our setting differs from the setting studied in prior works on spurious correlations (Sagawa et al., 2020), which train a model only on source data $\mathcal{D}_S$ and evaluate the model's performance on the hardest target distribution (i.e., worst-group accuracy). This is also different from the setting used in fine-tuning methods for zero-shot generalization (Wortsman et al., 2022; Kumar et al., 2022): such methods fine-tune a pretrained model on source data $\mathcal{D}_S$ and directly evaluate performance on target data $\mathcal{D}_T^i$ without any exposure to labeled target data. Compared to these zero-shot evaluation settings, we argue that a small amount of target data may realistically be required to handle the arbitrary distribution shifts that arise in the real world. Target data can be an effective point of leverage because it can be available or easy to collect, and we find that even a small dataset can reveal a lot about what features are effective in the target distribution. Our problem setting of adapting with target data has been used in some recent works (Kirichenko et al., 2022; Rosenfeld et al., 2022; Izmailov et al., 2022; Lee et al., 2022a), but we specifically focus on the setting in which we only have access to a very small target dataset, i.e., $M \ll N$.

C Proofs for Section 5

We present proofs for our theoretical analysis in Section 5 along with some additional statements. As in the main paper, we denote d as the dimensionality of the feature-space basis learned by Pro^2, D as the original dimension of the representations given by the feature backbone f, p_S as the source distribution, p_T as a target distribution, N as the number of source datapoints, and M as the number of target datapoints. We let $\mathbf{\Pi}_d$ denote the projection matrix for $\mathrm{span}(\{\Pi_i\}_{i=1}^d)$, $\mathbf{\Pi}_d$=$[\Pi_1, .., \Pi_d][\Pi_1, .., \Pi_d]^\top$. If the target error for the feature w is $\mathcal{L}_T(\mathbf{w}) := \mathbb{E}_{\mathcal{D}_T} l(\langle \mathbf{w}, \mathbf{x}\rangle,\ \mathrm{y})$, then the bias incurred by probing on the subspace $\mathbf{\Pi}_d$ consisting of source features is:

$$b_d \ := \min_{\mathbf{w}' \in \mathrm{span}(\{\Pi_i\}_{i=1}^d)} \mathcal{L}_T(\mathbf{w}') \ - \ \min_{\mathbf{w} \in \mathcal{W}} \mathcal{L}_T(\mathbf{w}),$$

and we denote the feature-space basis of dimensionality d learned by Pro^2 as follows:

$$\hat{\mathbf{w}}_d \triangleq \min_{\mathbf{w} \in \mathrm{span}(\{\Pi_i\}_{i=1}^d)} \sum_{i=1}^{M} l(\langle \mathbf{w}, \mathbf{x}^{(i)}\rangle,\ \mathrm{y}^{(i)}) \tag{3}$$

Theorem 4 (Information in projected input, Theorem 1)**.** *When the distributions $p((\mathbf{x} - \mathbb{E}[\mathbf{x}]) \mid \mathrm{y})$ are identical for each y. the solution $\{\Pi_i\}_{i=1}^d$ returned by* Pro^2 *maximizes the mutual information $I(\mathbf{A}\mathbf{x};\ \mathrm{y})$ (and a strict upper bound on it otherwise) among all $D \times d$ matrices $\mathbf{A}$ with orthogonal columns.*

Proof. We use an inductive argument on d. Consider the solution to

$$\max_{A \in B^{D \times d}} I(Ax; y).$$

If $d = 1$, then this returns the Bayes-optimal mean predictor, because the Bayes error for $p(y \mid w^\top x)$ is upper and lower bounded by Eq. 16 in Petridis and Perantonis (2004). If $d \geq 2$,

$$\max_{A \in B^{D \times d}} I(Ax; y) = \max_{A' \in B^{D \times (d-1)}} I\left(A'x, v^\top x; y\right), \text{where } v \in \text{Null}(A'), \|v\|_2 = 1.$$

Decomposing the right expression, we have

$$I\left(Ax, v^\top x; y\right) = I\left(A'x; y\right) + I\left(v^\top x; y\right) - I(v^\top x; Bx) + I(v^\top x; Bx \mid y) = I\left(A'x; y\right) + I\left(v^\top x; y\right),$$

because $I(v^\top x; Bx) = I(v^\top x; Bx \mid y)$ due to the assumption that the mean-centered distributions $(x - E[x])|y$ are identical. Thus, we have for $d - 1$:

$$\max_A I(Ax; y) = \max_{A' \in B^{D \times (d-1)}, v \in \text{Null}(A')} I\left(A'x; y\right) + I\left(v^\top x; y\right) \tag{4}$$

$$= \max_{A' \in B^{D \times (d-1)}} \left[I\left(A'x; y\right) + \max_{v \in \text{null}(A')} I\left(v^\top x; y\right) \right]. \tag{5}$$

With an inductive argument, for d, we have:

$$\max_{A \in R^{D \times d}} I(Ax; y) = \max_{A' \in \mathbb{R}^{D \times (d-1)}} I\left(A'x; y\right) + \max_{v \in \mathbb{R}^n} I\left(v^\top \left(I - A'A'^\top\right) x; y\right). \tag{6}$$

Applying iteratively, we have

$$\begin{aligned}
\max_{A \in \mathbb{R}^{D \times d}} I(Ax; y) = & \max_{v_1 \in \mathbb{R}^n} \left(v_1^\top x; y\right) \\
& + \max_{v_2 \in \mathbb{R}^n} I\left(v_2^\top \left(I - v_1^* {v_1^*}^\top\right) x; y\right) \\
& + \max_{v_3 \in \mathbb{R}^n} I\left(v_3^\top \left(I - v_2^* {v_2^\star}^\top\right) \left(I - v_1^\star {v_1^\star}^\top\right) x; y\right) + \dots,
\end{aligned}$$

where $v_1^\star, v_2^\star, \dots, v_d^*$ denote the solutions to each subsequent max term, and assume each term is unique. Now this sequence of solutions is the same as that returned by solving the following optimization problem iteratively:

1. $\min_{\|v\| \leq 1} l(\langle v, x \rangle, y)$
2. Project data onto $(I - vv^\top)x$
3. Re-solve (1.) to get next v and so on.

Finally, we claim that solution returned by this iterative optimization is the same as that returned by optimizing the projection of PRO^2. Our objective

$$\begin{aligned}
\min_{\Pi_1 \dots \Pi_d, \Pi_i \perp \Pi_j (i \neq j)} \sum_i l(\Pi_i x; y) &= \max_{\Pi_1 \dots \Pi_d} \sum_i I(\Pi; x; y) \\
&= \max_{\Pi_i} \left(\max_{\Pi_2 \dots \Pi_d} I(\Pi_1 x; y) + \sum_{j=2}^{d} I(\Pi_j (I - \Pi_1 \Pi_1^\top) x; y) \right)
\end{aligned}$$

Then, again using Eq. (16) in Petridis and Perantonis (2004) connecting cross entropy loss to Bayes error, the above is equivalent to (6), concluding our argument. □

Lemma 5 (bias induced by shift)**.** *For some $\mathbf{w}_T^*$ that is the Bayes optimal linear predictor on distribution p_T over the full feature space, and an L-Lipschitz smooth convex loss l, the bias $b_d \leq L \cdot \|(\mathbf{I}_D - \mathbf{\Pi}_d)\mathbf{w}_T^*\|_2$. When $\mathbf{\Pi}_d$ is a random rank d projection matrix with columns drawn uniformly over the sphere S^{d-1}, then $b_d L \sqrt{1 - \frac{d}{D}} \cdot \|\mathbf{w}_T^*\|_2$.*

Proof. Let l be L-Lipchitz, smooth, and convex. We have that the bias

$$
\begin{aligned}
b_d &= \min_{w \in \operatorname{span}\{\Pi_i\}_{i=1}^{d}} \mathbb{E} l(\langle w, x\rangle, y) - \min_{w \in \mathcal{W}} \mathbb{E} l(\langle w, x\rangle, y) \\
&= \min_{w \in \operatorname{span}\{\Pi_i\}_{i=1}^{d}} \mathbb{E}_{D_T} l(\langle w, x\rangle, y) - \min_{w \in \operatorname{span}\{\Pi_i\}_{i=1}^{D}} \mathbb{E}_{D_T} l(\langle w, x\rangle, y) \\
&= \min_{w \in \operatorname{span}\{\Pi_i\}_{i=1}^{d}} \mathbb{E}_{D_T} l(\langle w, x\rangle, y) - E_{D_T} l\left(\langle w_T^\star, x\rangle, y\right), \text{where} \quad w_T^\star = \min_{w \in \mathcal{W}} \mathbb{E} l(\langle w, x\rangle, y) \\
&= \min_{w \in \mathcal{W}} \mathbb{E}_{D_T} l\left(\left\langle w, \Pi_d \Pi_d^\top x\right\rangle, y\right) - \mathbb{E}_{D_T} l\left(\langle w_T^\star, x\rangle, y\right) \\
&\leq \mathbb{E}_{D_T} l\left(\left\langle w_T^*, \Pi_d \Pi_d^\top x\right\rangle, y\right) - \mathbb{E}_{D_T} l\left(\left\langle w_T^\star, \Pi_d \Pi_d^\top x + (I - \Pi_d \Pi_d^\top x\right\rangle, y\right) \\
&= \mathbb{E}_{D_T} l\left(\left\langle x, \Pi_d \Pi_d^\top w_T^\star\right\rangle, y\right) - l\left(\left\langle\left(I - \Pi_d \Pi_d^\top\right) w_T^\star + \Pi_d \Pi_d^\top w_T^\star, x\right\rangle, y\right). \qquad (7)
\end{aligned}
$$

Now let $f = l(\langle \cdot, x\rangle, y)$. Then

$$f(a) - f(a+b) \leq -b^\top \nabla_{a+b} f(a+b),$$

and Eq. 7 is convex in its first argument. If

$$a = \Pi_d \Pi_d^\top W_T^\star, a + b = \left(I - \Pi_d \Pi_d^\top\right) w_T^\star + \Pi_d \Pi_d^\top w_T^\star,$$

then we have

$$\text{Eq. (7)} \leq \left\| \left(I - \Pi_d \Pi_d^\top\right) w_T^\star \right\|_2 \left\| \nabla l(\cdot) \right\|_2.$$

$\left\| \nabla l(\cdot) \right\|_2 \leq L$(Lipschitz) and also ∇l exists everywhere (because it is smooth). Thus,

$$\text{Eq. (7)} \leq L \left\| \left(I - \Pi_d \Pi_d^\top\right) \omega_T^\star \right\|_2.$$

Let us consider a special case where $\mathbf{\Pi}_d$ is a random projection matrix. Thus, $\mathbf{I}_D - \mathbf{\Pi}_d$ is also a random $D - d$ projection matrix. Using standard high dimensional probability bounds for $|(\mathbf{w}_T^*)\top \mathbf{u}|$ for random vectors $\mathbf{u}$ drawn uniformly from S^{D-1} (refer Ch.3 in Wainwright (2019)), we get that

$$|(\mathbf{w}_T^*)\top \mathbf{u}| \in (\sqrt{1/D} \pm \sqrt{\log(1/\delta)/D})$$

with probability $\geq 1 - \delta$.

Applying this result to random $D - d$ projection $L \left\| \left(I - \Pi_d \Pi_d^\top\right) \omega_T^\star \right\|_2$ we get:

$$b_d L \sqrt{1 - (d/D)} \|\omega_T^\star\|_2$$

Now our proof is complete. □

Lemma 6 (generalization error). *For an L-Lipshitz, B-bounded loss l, with probability $\geq 1 - \delta$, $\hat{\mathbf{w}}_d$ in equation 3 has generalization error $\frac{\sqrt{d} + B\sqrt{\log(1/\delta)}}{\sqrt{M}}$, when $\|\mathbf{x}\|_\infty = O(1)$.*

Proof. For this proof, we use the following two statements.

Lemma 1 (Bartlett and Mendelson (2002)). *For an L-Lipshitz B-bounded loss l, the generalization error for predictor $\hat{\mathbf{w}}_d$, contained in the class of l_2 norm bounded linear predictors $\mathcal{W}$ is bounded with probability $\geq 1 - \delta$:*

$$l(\langle \hat{\mathbf{w}}_d, \mathbf{x}\rangle, \mathrm{y}) \; - \; \sum_{i=1}^{M} l(\langle \mathbf{w}, \mathbf{\Pi}_d \mathbf{x}^{(i)}\rangle, \; \mathrm{y}^{(i)}) \; \leq \; 2L\mathcal{R}_n(\mathcal{W}) + B\sqrt{\frac{\log(1/\delta)}{2M}}$$

where $\mathcal{R}_n(\mathcal{W})$ is the empirical Rademacher complexity of l_2 norm bounded linear predictors.

Lemma 2 ($\mathcal{R}_n(\mathcal{W})$ bound for linear functions Kakade et al. (2008)). *Let $\mathcal{W}$ be a convex set inducing the set of linear functions $\mathcal{F}(\mathcal{W}) \triangleq \{\langle \mathbf{w}, \mathbf{x}\rangle : \mathcal{X} \mapsto | \; w \in \mathcal{W}\}$ for some input space $\mathcal{X}$, bounded in norm $\|\cdot\|$ by some value $R > 0$. Now, if $\exists$ a mapping $h : \mathcal{W} \mapsto$ that is κ-strongly convex with respect to the dual norm $\|\cdot\|_*$ and some subset $\mathcal{W}' \subseteq \mathcal{W}$ takes bounded values of $h(\cdot)$ $\{h(\mathbf{w}) \leq K \mid \mathbf{w} \in \mathcal{W}'\}$ for some $K > 0$, then the empirical Rademacher complexity of the subset $\mathcal{W}'$ given by $\mathcal{R}_n(\mathcal{F}(\mathcal{W}')) \leq R\sqrt{\frac{2K}{\kappa n}}$.*

Let $\|\cdot\|_2^2$ be the function $h : \mathcal{W} \mapsto$ in Lemma 2, and we know that $\|\cdot\|_2^2$ is 2-strongly convex in l_2 norm. Further, take the standard l_2 norm as the norm over $\mathcal{X}$. So, the dual norm $\|\cdot\|_*$ is also given by l_2 norm. Thus, $\kappa = 2$. We also know that $\mathcal{W}$ is bounded in $\|\cdot\|_2$ by 1, based on our setup definition. Thus, $K = 1$.

Further, we note that $\|\mathbf{x}\|_\infty = O(1)$, thus apply Cauchy-Schwartz and using the fact that $\mathbf{\Pi}_d = 1$:

$$\|\mathbf{\Pi}_d\mathbf{x}\| \leq \mathbf{\Pi}_d\|\mathbf{x}\|_2$$
$$\|\mathbf{x}\|_2 \leq \sqrt{d}\|\mathbf{x}\|_\infty\sqrt{d}$$

Hence, $R\sqrt{d}$. Plugging this in to Lemma 2 we get the empirical Rademacher complexity $\mathcal{R}_M(\mathcal{W})\sqrt{d/M}$, and plugging this into Lemma 1 yields the main result in Lemma 6.

□

Theorem 7 (bias-variance tradeoff, Theorem 2). *When the conditions in Lemma 5 hold and when $\|\mathbf{x}\|_\infty = \mathcal{O}(1)$, for B-bounded loss l, w.h.p. $1-\delta$, the excess risk for the solution $\hat{\mathbf{w}}_d$ of* PRO^2 *that uses d features is $\mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w})$*

$$\|(\mathbf{I}_D - \mathbf{\Pi}_d)\mathbf{w}_T^*\|_2 + \frac{\sqrt{d} + B\sqrt{\log(1/\delta)}}{\sqrt{M}}, \tag{8}$$

where the first term controls the bias and the second controls the variance.

Proof. The excess risk for $\hat{\mathbf{w}}_d$ is given by: $\mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w})$.

$$\begin{aligned}
&\mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w}) \\
&= \mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\text{span}\{\Pi_i\}_{i=1}^d} \mathcal{L}_T(\mathbf{w}) + \min_{\mathbf{w}\in\text{span}\{\Pi_i\}_{i=1}^d} \mathcal{L}_T(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w}) \\
&\min_{\mathbf{w}\in\text{span}\{\Pi_i\}_{i=1}^d} \mathcal{L}_T(\mathbf{w}) - \min_{\mathbf{w}\in\mathcal{W}} \mathcal{L}_T(\mathbf{w}) + \mathcal{L}_T(\hat{\mathbf{w}}_d) - \min_{\mathbf{w}\in\text{span}\{\Pi_i\}_{i=1}^d} \mathcal{L}_T(\mathbf{w}) \\
&\|(\mathbf{I}_D - \mathbf{\Pi}_d)\mathbf{w}_T^*\|_2 + \frac{\sqrt{d} + B\sqrt{\log(1/\delta)}}{\sqrt{M}}
\end{aligned}$$

where the first term is the bias (bounded using Lemma 5), and the second term is the generalization error or the variance (bounded using Lemma 6). □

Corollary 8. *Under the SHOG model, Π_1 recovers the linear discriminant analysis (LDA) solution, $\Pi_1 = \Sigma^{-1}(\mu_2 - \mu_1)/(\|\Sigma^{-1}(\mu_2 - \mu_1)\|_2)$.*

Proof. Since LDA solution is Bayes optimal under the HOG model, it is exactly characterized by the top eigen vector of $\Sigma^{-1}(\mu_2 - \mu_1)(\mu_2 - \mu_1)^\top$. Thus, the Bayes optimal solution on target $\mathbf{w}_T^* \propto \Sigma^{-1}(\mu_2 - \mu_1)$, and since Π_1 returns the Bayes optimal linear predictor, following Theorem 1, the above corollary is proven. □

Lemma 9 (bias under SHOG). *When $\mathbf{\Pi}_d$ is returned by* PRO^2*, the bias b_d term under our SHOG is $b_d\|(\mathbf{I}_D - \mathbf{v}_S\mathbf{v}_S^\top)\mathbf{v}_T\|$. Here, $\mathbf{v}_S = \frac{\Sigma_S^{-1}\mu}{\|\Sigma_S^{-1}\mu\|_2}$ and $\mathbf{v}_T = \frac{\Sigma_T^{-1}\mu}{\|\Sigma_T^{-1}\mu\|_2}$. Further, when $\|\Sigma_S\|_{\text{op}}$ is bounded, and $\mathbf{\Pi}_d$ is a random rank d projection matrix, $b_d = \mathcal{O}\sqrt{1 - \frac{d}{D}} \cdot \text{KL}(p_S||p_T)$.*

Proof. Since,

$$b_d \leq \|(\mathbf{I}_D - \mathbf{\Pi}_d)\mathbf{w}_T^*\|_2 \leq |(\mathbf{I}_D - \mathbf{\Pi}_1)\mathbf{w}_T^*\|_2.$$

From corollary 8, we know that $\mathbf{\Pi}_1$ is exactly the rank-1 projection matrix given by the direction $\Sigma_S^{-1}(\mu_2 - \mu_1)/(\|\Sigma_S^{-1}(\mu_2 - \mu_1)\|_2)$. This gives us the first result for $\mathbf{v}_S, \mathbf{v}_T$.

For the second result we rely on the convexity of KL divergence and KL divergence for multivariate Gaussian distributions to get:

$$\begin{aligned}
&\mathrm{KL}(p_S||p_T) = \mathrm{KL}(p(y)p_S(\mathbf{x} \mid \ ry)||p(y)p_T(\mathbf{x} \mid \ ry)) \\
&\leq \mathrm{KL}(p_S(\mathbf{x} \mid \ ry)||p_T(\mathbf{x} \mid \ ry)) \\
&= 0.5 \cdot \mathrm{KL}(\mathcal{N}(\mu_1, \Sigma_S)||\mathcal{N}(\mu_1, \Sigma_T)) + 0.5 \cdot \mathrm{KL}(\mathcal{N}(\mu_2, \Sigma_S)||\mathcal{N}(\mu_2, \Sigma_T)) \\
=&\frac{1}{2}\mathrm{tr}(\Sigma_T^{-1}\Sigma_S) - \sum_{i=1}^{D} \log \lambda_i^S + \sum_{i=1}^{D} \log \lambda_i^T - D
\end{aligned} \tag{9}$$

Refer to Wainwright (2019) for the final step, where λ_i^S and λ_i^T are the eigen values of source target covariances.

The final term in the above derivation is $\mathcal{O}(\mathrm{tr}(\Sigma_T^{-1}))$ when $\Sigma_S = O(1)$.

From Lemma 5 we know that under random projections onto d dimensions,

$$b_d \ \leq \ L \cdot \sqrt{1-(d/D)} \|\mathrm{w}_T^*\| \cdot \sqrt{1-(d/D)} \|\Sigma_T^{-1}(\mu_2 - \mu_1)\| \mathrm{tr}(\Sigma_T^{-1}) \tag{10}$$

where we use Corollary 8. Thus from equation 10, equation 9, we get our desired bound:

$$b_d \sqrt{1 - \frac{d}{D}} \cdot \mathrm{KL}(p_S||p_T)$$

□

Corollary 10 (tradeoff under SHOG, Corollary 3)**.** *Under our SHOG model of shift, and conditions for a random projection* $\mathbf{\Pi}_d$ *in Lemma 9, the target error* $\mathcal{L}_T(\hat{\mathbf{w}}_d)\mathcal{O}\sqrt{1 - \frac{d}{D}} \cdot \mathrm{KL}(p_S||p_T) + \sqrt{\frac{d}{M}}$*, when* $\Sigma_T = O(1)$.

Proof. Direct application of the variance result in Lemma 6 and bias result in Lemma 9, using the same technique used to prove Theorem 2. □

D EXPERIMENTAL DETAILS

D.1 PYTORCH PSEUDOCODE FOR THE PROJECTION STEP OF PRO^2

Below, we provide PyTorch pseudocode for the projection step of PRO^2 for binary classification datasets.

```
def learn_feature_space_basis(x, y, num_features):
    projection = torch.nn.Linear(x.shape[1], num_features)
    opt = torch.optim.AdamW(projection.parameters(), lr=0.01,
                                             weight_decay=0.01)
    max_steps = 100
    for i in range(max_steps):
        logits = projection(x)
        loss = F.binary_cross_entropy_with_logits(logits, y, reduction="
                                                 none").mean()
        opt.zero_grad()
        loss.backward()
        opt.step()
        # Enforce orthogonality; we're performing projected gradient
                                                 descent
        Q, R = torch.linalg.qr(linear_model.weight.detach().T)
        projection.weight.data = (Q * torch.diag(R)).T
    feature_space = projection.weight.detach().T
    return feature_space
```

D.2 Additional dataset details

- **4-Way Collages** (Teney et al., 2021). This binary classification dataset consists of 4-way collages of four images per datapoint, one from each of (1) CIFAR, (2) MNIST, (3) Fashion-MNIST, and (4) SVHN. All four image features are completely correlated in the source data, and we consider four target distributions, where only one of the image features are predictive of the label in each target distribution.
- **Waterbirds** (Sagawa et al., 2020). This dataset tasks the model with classifying images of birds as either a waterbird or landbird. The label is spurious correlated with the background of the image, which is either water or land. There are 4,795 training samples, of which 95% of the data follows the spurious correlation. We use the original training set as the source data and evaluate on 3 different target distributions constructed from the original test dataset: (1) Minority, which contains the test data points that do not follow the spurious correlation, (2) Spurious, containing the points that do, and (3) Balanced, which contains an equal number of points from each of the 4 (bird, background) groups.
- **CelebA** (Liu et al., 2015). Similar to Waterbirds, we use the original training set as source data and evaluate on (1) Minority, (2) Spurious, and (3) Balanced target distributions. In our main experiments in Section 6, we use target distributions corresponding to the spurious correlation typically used for evaluation (spurious attribute–gender with label–hair color). Below, in Appendix E include additional results on 4 other variants following the settings used in Lee et al. (2022b): (1) CelebA-1 uses slightly open mouth as the label and wearing lipstick as the spurious attribute, (2) CelebA-2 uses attractive as the label and smiling as the spurious attribute, (3) CelebA-3 uses wavy hair as the label and high cheekbones as the spurious attribute, and finally (4) CelebA-4 uses heavy makeup as the label and big lips as the spurious attribute.
- **Camelyon17** (Bandi et al., 2018). This dataset is part of the WILDS benchmark Koh et al. (2021) and contains medical images where variations in data collection from different hospitals induce naturally occurring distribution shifts. We evaluate on 2 target distributions: (1) ID-Test: a held out test set of images from the source distribution, and (2) OOD-Test: the actual test distribution with a distribution shift due to evaluating data from a different hospital.

Pre-trained models and additional training details. We extract penultimate embeddings of all source and target datapoints from a pre-trained backbone. We preprocess all datapoints according to the augmentation used during pre-training, and obtain feature embeddings with eval-mode batch normalization. We cache all embeddings for a (backbone, dataset) pair to a single file and train our linear models from the cached file. We use CLIP-ViT-L/16 Dosovitskiy et al. (2020) in our main experiments, and additionally experiment with ResNet18 He et al. (2016), ResNet50, ResNet50-SWaV Caron et al. (2020), CLIP-ViT-B/16 models in Appendix E.5. All pretrained models are publicly available online. We train all models using the AdamW optimizer Loshchilov and Hutter (2017) with weight decay 0.01. For all experiments, we perform early stopping with accuracy on held-out target data and report mean and standard deviation across 10 runs.

E Additional Experimental Results

E.1 Additional visualizations for synthetic Gaussian experiment

In Figure 5, we approximate the bias and variance in the synthetic HOG experiment studied in Figure 2. On the left, for each test distribution (ID, Near OOD, and Far OOD), we plot the relationship between approximate bias (using error at the largest target dataset size) and nullspace norm and find that they have a roughly linear relationship. Thus, this plot empirically supports the connection supported in the theory between bias and the number of features used, as the nullspace norm decreases as the dimension of the feature-space basis increases. On the right, Hence, we connect

E.2 Are causal features always best?

In this experiment, we aim to demonstrate how the different features that are predictive on source data can perform differently on different target distributions. On the Waterbirds dataset, we learn

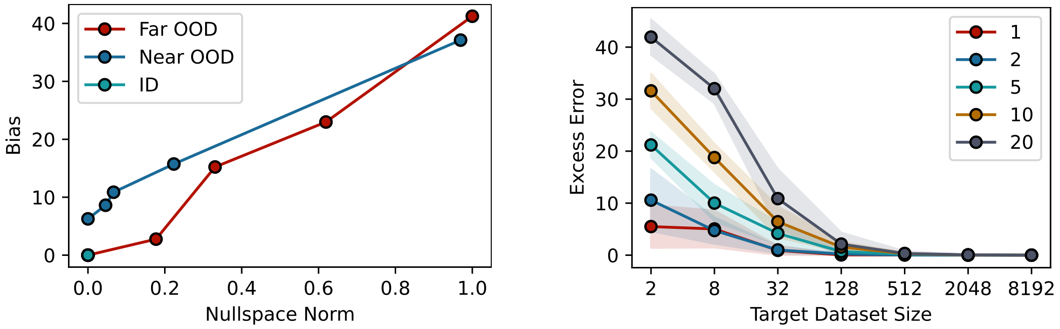


Figure 5: Visualization of bias and variance in the synthetic homoscedastic Gaussian experiment Figure 2. (Left) We approximate bias by the error at the largest target dataset size, and compare to the nullspace norm. The two quantities have a roughly linear relationship. (Right) We approximate variance by the difference between the error at each dataset size and the error at the largest. We report the average across the three test distributions. Note on the left plot, ID is easily learned and so the corresponding line is therefore clustered near (0, 0), as the nullspace norm and bias are both near 0.

	Majority Groups		Minority Groups	
	LB+L	WB+W	LB+W	WB+L
Causal	93.6 (1.1)	95.1 (0.5)	**90.3 (0.5)**	**94.3 (0.4)**
Shortcut	**96.5 (0.8)**	**98.0 (0.4)**	38.3 (4.1)	91.2 (1.1)

Table 1: Different features can be best for different target distributions. We learn two linear classifiers for Waterbirds based on the causal and shortcut features, respectively. We report average accuracy within each group, and show standard deviation inside parentheses. LB and WB represent landbirds and waterbirds, and L and W represent land and water backgrounds. While the causal feature achieves higher worst-group accuracy, the shortcut feature achieves higher accuracy on the majority groups.

two linear classifiers on top of backbone embeddings. We learn an oracle feature by minimizing worst-group loss (Group DRO, Sagawa et al. (2020)), and an oracle shortcut classifier by minimizing average loss on the majority data. These are the same features used for Figure 6. In Table 1, as expected, the causal feature achieves the best worst-group accuracy. However, we find that the shortcut feature outperforms the causal feature on the two majority groups, indicating that this feature would achieve higher performance in a distribution skewed towards majority groups. In particular, such shortcut features are especially useful on certain distributions when fairness metrics do not matter, e.g. like positions of cars. In other words, there is no one best feature, and different features can be best for different target distributions. These observations motivate PRO^2: it can be beneficial to extract a diverse set of features that cover both causal and shortcut features, and adapt to different target distributions by interpolating between these learned features.

E.3 Empirical analysis of projected feature space

We begin by observing the empirical properties of the projected feature space learned during the first projection phase of PRO^2. The Waterbirds dataset consists of "spurious" groups where the background type (land or water) correlates with the bird type (land or water), on which using a shortcut feature that relies on background type will perform optimally, as well as "minority" groups in which the correlation does not hold and requires a robust feature that focuses on the bird itself. On this dataset, we first extract oracle shortcut and robust features by minimizing loss on spurious and minority groups on target data, respectively. These two directions serve as proxies for the optimal classifier on two different target distributions. In addition to PRO^2, we also evaluate a random feature extraction method, which simply samples a random orthonormal basis for the original ^{D} embedding space. We plot the nullspace norm of these two features in the subspace spanned by the first k directions, for $1 \leq k \leq D = 1024$ in Figure 6. As expected, we see that the earlier features learned by PRO^2 are more similar to the shortcut feature than the robust feature. Because the orthogonality

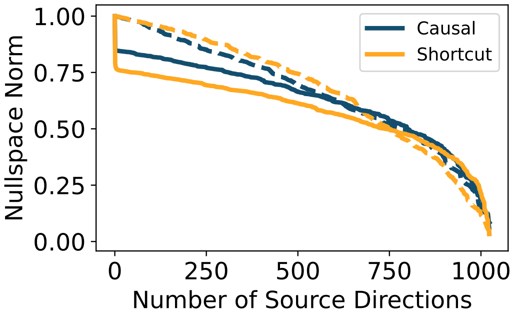


Figure 6: **Nullspace norm on Waterbirds.** We plot the nullspace norm (y-axis) of the shortcut and causal features in the subspace spanned by the first d directions (x-axis) learned by PRO^2 (solid lines) or a random orthonormal basis (dotted lines). We find that compared to random, the first few features learned by PRO^2 are informative and therefore have lower nullspace norm. Additionally, by enforcing orthogonality, the features learned eventually fully cover both types of features, with the nullspace reducing to zero.

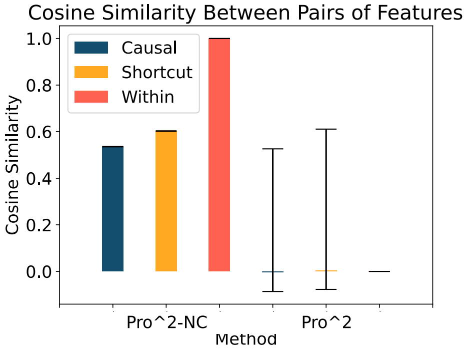


Figure 7: We show the average cosine similarity between randomly chosen pairs of individual features taken from features learned by the method PRO^2-NC and PRO^2 with causal and shortcut features on the Waterbirds dataset. The error bars are the minimum and maximum cosine similarity for pairs of features from the corresponding methods. In contrast to PRO^2-NC, the features learned by PRO^2 have very little similarity with each other, although the max similarity between a features learned by PRO^2 and both the shortcut and causal features is still high, allowing PRO^2 to cover a more diverse range of features.

constraint forces the features to be different from each other, the nullspace norm reduces to zero at the highest value $k = 1024$. This experiment shows that the basis learned by PRO^2 contains both the robust and shortcut features for this dataset, and that the robust and shortcut features emerge even for very low-rank bases (i.e., for small values of d). In contrast, a random orthogonal basis only captures these two predictive features when the rank is larger. This indicates that our orthogonal projection approach quickly picks up on the most important directions in feature space, which in this case correspond to the shortcut feature representing the background and the robust feature representing the type of bird, as discussed in prior work (Sagawa et al., 2020).

E.4 FEATURE SIMILARITY

We also compare PRO^2 and PRO^2-NC to see how the orthogonality constraint effects feature diversity In Figure 7, we plot the average cosine similarity between the shortcut and causal features with two versions of PRO^2: one with no constraints and one with orthogonality enforced. More specifically, for each bar, we calculate the average cosine similarity between 200 randomly chosen features learned by the method (either PRO^2-NC or PRO^2) with the causal and shortcut features learned above along with another randomly chosen feature from the method (labeled "Within"). The error bars are the minimum and maximum cosine similarity for a pair of features from the corresponding methods.

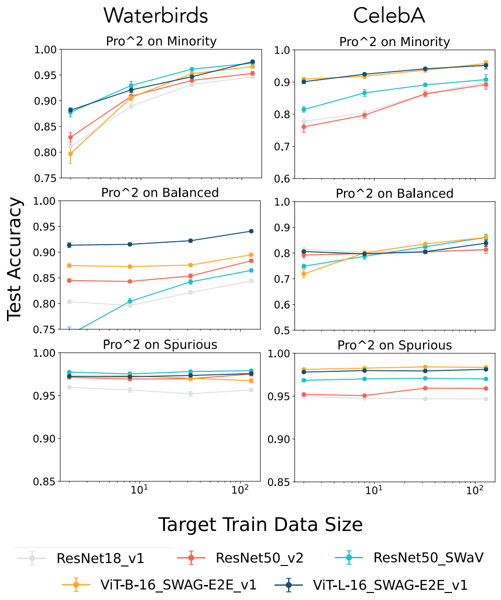


Figure 8: **Different backbones.** We show the accuracy of PRO^2, where we use various pretrained backbones, which are not fine-tuned. PRO^2 is able to leverage improvements in the backbone with minimal computational overhead.

From this plot, we see that when orthogonality is not enforced, the features learned are not diverse: with PRO^2-NC, the “Within” column has high cosine similarity and very little variation, showing that all features are very similar to each other, and they are all more similar to the shortcut feature than the causal feature. Thus, interpolating between such features may struggle to adapt to target distributions that require reliance on the causal feature. On the other hand, the features learned by PRO^2 have very little similarity with each other, although the max similarity between a features learned by PRO^2 and both the shortcut and causal features is still high. Thus, enforcing orthogonality is important for learning diverse features that span both the shortcut and causal features.

E.5 USING VARIOUS PRETRAINED BACKBONES

Finally, as PRO^2 relies on using a pre-trained backbone model that is not fine-tuned to initially extract features, we study how different backbones affect performance. In Figure 8, we plot the accuracy of PRO^2 using 5 pre-trained backbone models that achieve a range of Image-Net accuracies. We find that PRO^2 improves significantly with better pre-trained backbones. These experiments demonstrate the promise of the PRO^2 framework. The quality of pre-trained feature extractors will continue to improve with future datasets and architectures, and PRO^2 leverages such pre-trained backbone models for distribution-shift adaptation in a computationally efficient manner.

E.6 ABLATION ON THE IMPORTANCE OF ENFORCING ORTHOGONALITY

For the purposes of our empirical analysis, we additionally consider a simpler variant that optimizes the projection matrix Π with **N**o **C**onstraint on orthogonality:

$$\Pi_i = \arg\min \mathbb{E}_{(x,y)\sim\mathcal{D}_S}\mathcal{L}(\Pi_i(f(x)), y). \tag{PRO^2-NC}$$

We compare PRO^2 to PRO^2-NC in Figure 9. While PRO^2-NC is is sufficient in some scenarios with milder distribution shift, where the shortcut feature continues to be informative, it fails to learn a diverse set of predictive features and often only learns shortcut features, often failing on more severe distribution shifts.

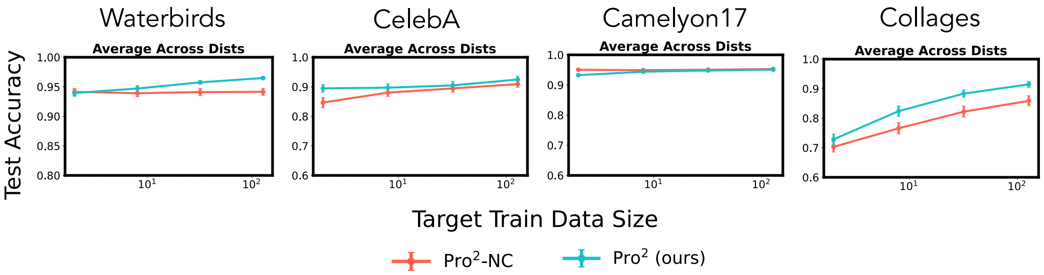


Figure 9: **Importance of orthogonality.** We show the adaptation accuracy of PRO^2 compared to PRO^2-NC, a variant without orthogonality enforced, averaged across the varying target distributions for each dataset.

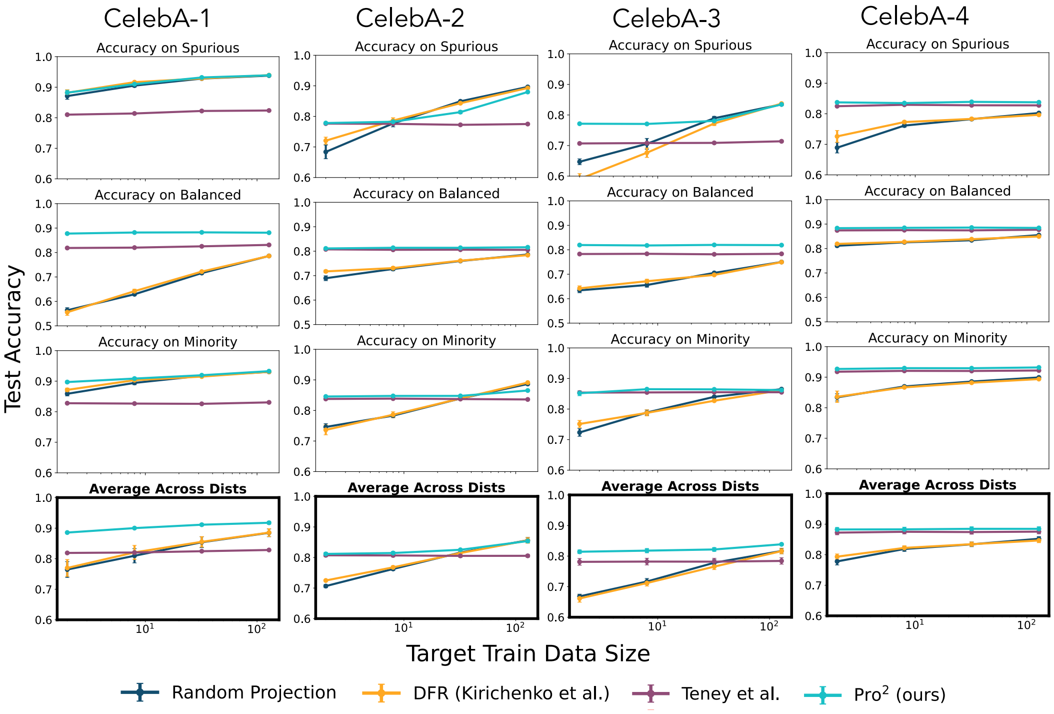


Figure 10: **Main results on CelebA variants.** We compare 4 different methods for learning features to adapt to a target distribution: (1) Random Projection, (2) DFR Kirichenko et al. (2022), i.e. standard linear probing, (3) Teney et al. (2021), and (4) PRO^2. We report target accuracies after probing with different target dataset sizes; we report mean and standard deviation across 10 runs. Similar to the trends seen in Figure 3, PRO^2 achieves high accuracy in the low-data regime, substantially outperforming both random orthogonal projection and no projection in most target distributions on all four datasets.

E.7 Evaluation on additional CelebA variants

Finally, in Figure 10 we supplement our main results in Figure 3 with additional results from 4 additional variants of CelebA. The takeaways from these results line up with those from Figure 3. In the few-shot adaptation problem setting, PRO^2 is consistently the most effective, compared to Random Projection, DFR Kirichenko et al. (2022), which uses standard linear probing, and Teney et al. (2021).