Distinguishing rule- and exemplar-based generalization in learning systems

Ishita Dasgupta [*1@] **Erin Grant** [*2] **Thomas L. Griffiths** [1]

Abstract

Machine learning systems often do not share the same inductive biases as humans and, as a result, extrapolate or generalize in ways that are inconsistent with our expectations. The trade-off between exemplar- and rule-based generalization has been studied extensively in cognitive psychology; in this work, we present a protocol inspired by these experimental approaches to probe the inductive biases that control this trade-off in category-learning systems. We isolate two such inductive biases: feature-level bias (differences in which features are more readily learned) and exemplar or rule bias (differences in how these learned features are used for generalization). We find that standard neural network models are feature-biased and exemplar-based, and discuss the implications of these findings for machine learning research on systematic generalization, fairness, and data augmentation.

1. Introduction

Extrapolation or generalization—decisions on unseen datapoints—is always underdetermined by data; which particular extrapolation behavior a machine learning (ML) system exhibits is determined by its inductive biases (Mitchell, 1980). When those inductive biases are opaque—as is often the case with many modern ML systems (Geirhos et al., 2020; D'Amour et al., 2020)—we can instead turn to empirical investigation of the *behavior* of a system to reveal the system's *implicit* inductive biases. Cognitive psychology provides a rich basis for experimental designs to study the often-opaque human cognitive system via its external behavior; these designs can be leveraged to distinguish between competing hypotheses about a machine learning system's inductive biases as well (*e.g.,* Ritter et al., 2017b; Lake et al., 2018; Dasgupta et al., 2019).

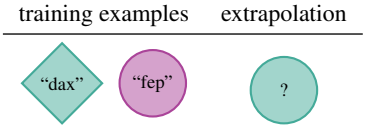


Figure 1. **Example of a data condition:** Data often underdetermines a decision boundary; here, it is unclear whether shape or color determines object label ("dax" vs "fep"). How a learner extrapolates to new stimuli reveals inductive bias.

We draw on cognitive psychology to construct a protocol that isolates the inductive biases determining how an ML system generalizes feature-based categories such as those in Fig. (**1**). A key property of such categorization problems is the presence of a *distractor* dimension that does not play a causal role in the underlying category boundary; the ground truth categorization is determined by a *discriminant* dimension. Such problems are ubiquitous in machine learning applications (*e.g.,* Beery et al., 2018), where learned associations between the distractor and the categorization label are termed "spurious" (Arjovsky et al., 2019). The tendency to acquire (potentially harmful) spurious associations is an example of a downstream consequence of implicit inductive bias, and so characterizing such implicit inductive biases is of both theoretical and practical interest.

We use abstract problem settings such as that in Fig. (**1**) to identify and isolate two distinct inductive biases underlying feature-based category learning. The first, *feature-level bias*, expresses a preference for some features over others to support a decision boundary (*e.g.,* preferring shape over color). The second, *exemplar bias*—vs. *rule bias*—expresses a preference for feature-dense (vs. feature-sparse) decision boundaries (*e.g.,* a boundary informed by both shape and color, vs. only one of the two features). Our protocol presents data conditions that manipulate feature co-occurrences observed during training such that the resulting extrapolation behavior is diagnostic of these inductive biases in the learner.

[*]Equal contribution [1]Departments of Psychology & Computer Science, Princeton University [2]Department of Electrical Engineering & Computer Sciences, UC Berkeley. [@]Now at DeepMind. Correspondence to: Ishita Dasgupta <idg@deepmind.com>, Erin Grant <eringrant@berkeley.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

Code at https://github.com/eringrant/icml-2022-rules-vs-exemplars.

The experimental setup underlying our training and testing conditions is similar to existing works in "combinatorial generalization" (Andreas et al., 2016; Johnson et al., 2017) and "subgroup fairness" (Sagawa et al., 2020a;b). Our work also makes several independent contributions: We identify and isolate two distinct inductive biases that affect extrapolation of feature-based categories, and we examine these across models in an expository points-in-a-plane setting, as well as in more naturalistic text and image domains. We demonstrate that existing measures of feature co-occurrence and extrapolation behavior ("spurious correlation" and "worst-group accuracy," Sagawa et al., 2020b) are insufficient to characterize these inductive biases. Finally, we consider the normative question: *What extrapolation behavior is desirable for a given application*? We provide a preliminary answer by discussing the relevance of the inductive biases we identify to related work in systematic generalization, fairness, and data augmentation.

2. Inductive biases in category learning

We start by introducing the two inductive biases of interest. **Feature-level bias** characterizes *which* feature a system finds *easier* or *harder* to learn and thus which feature a system will utilize when both are associated with the category label. This kind of feature-level bias has been studied extensively in human cognition (Landau et al., 1988; Hudson Kam & Newport, 2005), and specific feature-level biases—mostly notably the "shape-bias," the tendency to generalize image category labels according to shape rather than according to color or texture—have been revisited in the context of recent neural network models (Ritter et al., 2017a; Hermann et al., 2019; Geirhos et al., 2018). We examine feature-level bias for arbitrary features, as well as demonstrate how this bias interacts with—but is distinct from—another kind of inductive bias, to be discussed next.

Exemplar (or **rule**) **bias** characterizes *how* a system uses features to inform decisions by trading off *exemplar- and rule-based generalization*. A rule-based decision is made on the basis of minimal features that support the category boundary (*e.g.,* Ashby & Townsend, 1986), while an exemplar-based decision-maker generalizes on the basis of similarity to category exemplars (*e.g.,* Shepard & Chang, 1963), invoking many or all features that underlie a category. Extensive empirical work in cognitive psychology has found evidence of both kinds of generalization in humans (Nosofsky et al., 1989; Rips, 1989; Allen & Brooks, 1991; Smith & Sloman, 1994). This trade-off can be understood as a continuum that varies the number of features employed to discriminate between categories (Pothos, 2005).

Feature-level bias and exemplar bias are **practically relevant** because they describe how a learning system uses features to extrapolate, and different problem settings call for different ways of doing so. An exemplar-based system that depends on all features, and is not invariant to any of them, suffers when not all feature combinations are observed and systematic generalization to unobserved combinations is expected (Lake et al., 2018; Marcus, 2018; Arjovsky et al., 2019). On the other hand, a rule-based system that applies the same category decision rules across all data regions might over-generalize, which is undesirable in naturally occurring long-tailed distributions (Feldman & Zhang, 2020; Feldman, 2020; Brown et al., 2020). Diagnosing exemplar vs. rule bias is therefore of both theoretical and practical interest. In Section (**6**), we give a concrete example in a fairness setting—where certain regions of the data support is underrepresented but we want comparable accuracy on these regions nonetheless—in which understanding the inductive biases of the learning system allows for a data intervention that improves performance.

We now build intuitions for how **the category learning paradigm in Fig. (2) isolates feature-level bias and exemplar bias**. The stimuli Fig. (**2**) vary along two feature dimensions, shape and color. Color determines the label of an object (*i.e.,* green objects are "dax"; purple are "fep", using arbitrary names to emphasize that the category is novel to humans as well as to ML systems). Shape is unrelated to the underlying category structure and acts as a distractor. Participants (either humans or artificial learning systems) are independently placed in three different conditions—**cue conflict**, **zero shot**, or **partial exposure**—that vary in coverage of the feature space. After observing the *training examples*, the participant is presented with an *extrapolation* test consisting of an example outside the support of feature combinations observed during training (*i.e.,* they must classify the green circle as a "dax" or a "fep"). We explain below how differences in classification behavior on this extrapolation test isolate feature-level bias as well as exemplar-vs-rule bias, but first: We encourage the reader to try the experiment themselves to examine their intuitions.

Cue conflict (CC, top row, Fig. (**2**)). The data presented in this condition confound color and shape (*i.e.,* color and shape are equally predictive of the category boundary). How a system generalizes here directly measures its feature-level bias towards color or shape.

Characteristic behavior (right half of Fig. (**2**)). Since humans have an established shape bias (Landau et al., 1988), we expect that humans will classify the test item according to the object that shares its shape, not its color; in this case, as a "fep." However, this inductive bias is independent of whether a reasoner is rule- or exemplar-based; neither has an *a priori* propensity for features, both are equally likely to classify the test item as a "dax" or a "fep."

Zero shot (ZS, middle row, Fig. (**2**)). This condition requires extrapolation to a new feature value "zero-shot"

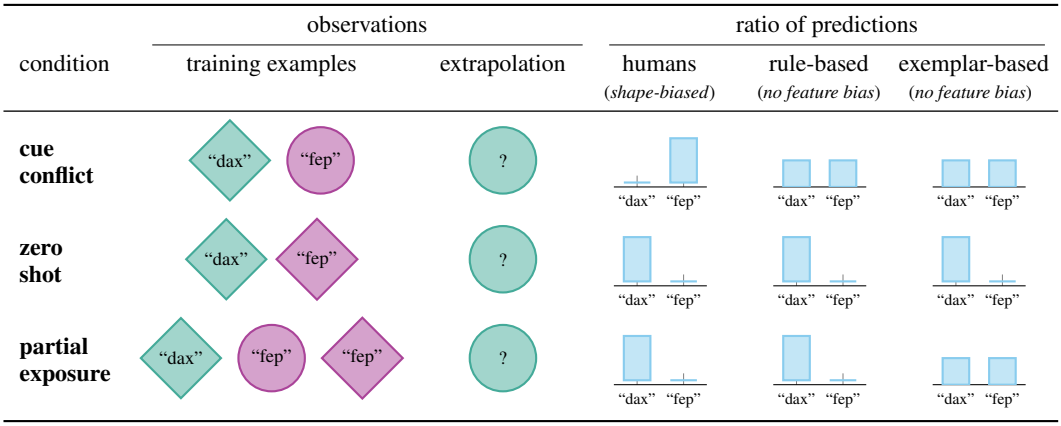


Figure 2. **Illustrative category learning experiment:** Training examples from the 3 independent training conditions, the extrapolation test, and characteristic behavior for learners with different inductive biases. We formalize the training conditions in Fig. (3).

(*i.e.,* without prior exposure). This setting is often used to examine out-of-domain (OOD) and compositional generalization in machine learning (Xian et al., 2018). Behavior in this condition reveals whether the model has learned the discriminating features and whether it can extrapolate to new feature values, and thus acts as a baseline.

Characteristic behavior (right half, Fig. (2)). Rule- and exemplar-based behavior in this condition is confounded. A rule-based learner infers the minimal rule that color determines label, does not assign any predictive value to shape, and therefore classifies the extrapolation stimulus based on color as a "dax." An exemplar-based learner categorizes based on the similarity along all feature dimensions of the extrapolation stimulus to category exemplars. Neither training exemplars have any overlap with the test stimulus along the shape dimension, but the "dax" overlaps along the color dimension, and the learner categorizes it as a "dax."

Partial exposure (PE, bottom row, Fig. (2)). Compared to zero shot, participants in this condition also receive "partial exposure" to a new feature value (*i.e., circle*) along the shape dimension. The extrapolation test in this condition is most similar to *combinatorial zero-shot generalization* (*e.g.,* Lake & Baroni, 2018a), where the learner is exposed independently to all feature values but has to generalize to a new combination.

Characteristic behavior (right half of Fig. (2)). This setting meaningfully distinguishes rule- and exemplar-based generalization. To understand this distinction, we contrast this condition to the cue-conflict condition. The addition of the purple diamond-shaped "fep" means the learner has seen both a diamond and a circle labeled "fep". A rule-based system takes this as direct evidence that shape is *not* predictive of category label and classifies the extrapolation stimulus on the basis of color as a "dax." This is typically also how humans extrapolate. This additional training example, however, does not impact an exemplar-based system, since it does not share any features with the extrapolation stimulus. The exemplar-based reasoner classifies on the basis of feature-overlap with training exemplars and is therefore indifferent, exactly as in the cue-conflict condition.

3. A protocol for measuring inductive bias

We embed the structure of the category learning problem discussed in Section (2) into a statistical learning problem that can be applied across domains to test black-box learners.

Problem setting. We consider the *oracle* compositional setting of Andreas (2019) in which inputs are a composition of categorical attributes with two latent binary features, $\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}}, \in \{0, 1\}$ that jointly determine the observation $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$ via some mapping $g : \{0, 1\}^2 \to \mathcal{X}$; see Fig. (3). We consider the binary classification task of fitting a model $\hat{f} : \mathcal{X} \to \{0, 1\}$ from a given model family $\mathcal{F}$ to predict a class for each observation. One of the underlying features, the *discriminant*, $\mathbf{z}_{\text{disc}}$, defines the decision boundary; the other one, the *distractor*, $\mathbf{z}_{\text{dist}}$, is not independently predictive of the label.

This specifies a generative process $\mathbf{x}, \mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}} \sim p(\mathbf{x} \mid \mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})\, p(\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$. $p(\mathbf{x} \mid \mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$ is either generated (*e.g.,* in Section (4)), or the empirical distribution of the subset of datapoints $\mathbf{x}$ with the corresponding underlying feature values (assuming access to these annotations, *e.g.,* in Sections (5) and (6)). $p(\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$ is varied across training conditions, as outlined below.

Training conditions. The upper-right quadrant in all subfigures of Fig. (3), for which $p(\mathbf{z}_{\text{disc}} = 1, \mathbf{z}_{\text{dist}} = 1) = 1$, acts as a hold-out set on which we can evaluate generalization to an unseen combination of attribute values. We produce multiple training conditions with the remaining three quadrants of data by manipulating $p(\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$. All the analyses

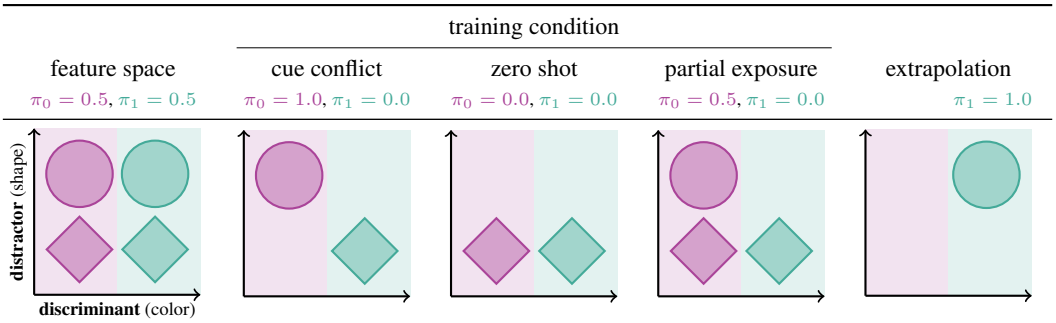


Figure 3. **Formalizing the illustrative experiment:** The experiment from Fig. (2) expressed in terms of the formalism in Section (3) with color as $\mathbf{z}_{\text{dist}}$ and shape as $\mathbf{z}_{\text{disc}}$. Background colors indicate the true category.

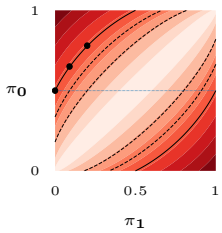


Figure 4. **Spurious correlation** (Eq. (3)).

in this work compare model extrapolation to the held-out test quadrant across various training conditions.

To equalize the class base rates we balance all training conditions across the discriminant; *i.e.,* we enforce $p(\mathbf{z}_{\text{disc}} = 0) = p(\mathbf{z}_{\text{disc}} = 1) = 0.5$. We also fix the number of datapoints across all conditions at N; With these constraints, we can control $p(\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$ via two degrees of freedom: $\pi_0 = p(\mathbf{z}_{\text{dist}} = 1 \mid \mathbf{z}_{\text{disc}} = 0)$ (this implicitly fixes $p(\mathbf{z}_{\text{dist}} = 0 \mid \mathbf{z}_{\text{disc}} = 0) = 1 - \pi_0$ to balance the dataset); and $\pi_1 = p(\mathbf{z}_{\text{dist}} = 1 \mid \mathbf{z}_{\text{disc}} = 1)$.The three conditions in Section (2), as well as the held-out test set, correspond to particular settings of π_0 and π_1 (shown in Fig. (3), more in Appendix (A.2)).

Measuring inductive bias. We measure *feature-level bias* as deviation from chance performance in the CC condition. *Exemplar bias* is measured as the difference between performance in the partial-exposure condition and zero-shot condition—-no difference indicates rule-based generalization, the magnitude of the difference measures exemplar bias. Formally, for a given model family $\mathcal{F}$, let $\hat{f}^{\text{ZS}}$ denote the result of selecting a model from $\mathcal{F}$ by training in the zero-shot condition, and similarly $\hat{f}^{\text{PE}}$ and $\hat{f}^{\text{CC}}$. We define feature-level bias (FLB) and exemplar vs. rule propensity (EvR) as:

$$\text{FLB}(\mathcal{F}) = \mathbb{E}[(A(y, \hat{f}^{\text{CC}}(\mathbf{x}))] - 0.5\,, \tag{1}$$

$$\text{EvR}(\mathcal{F}) = \mathbb{E}[A(y, \hat{f}^{\text{ZS}}(\mathbf{x}))] - \mathbb{E}[A(y, \hat{f}^{\text{PE}}(\mathbf{x}))] \tag{2}$$

where the expectation is taken with respect to the data distribution under the extrapolation region ($p(\mathbf{x}, y \mid \pi_0 = 1, \pi_1 = 1)$), and A is the 0-1 accuracy. FLB takes values between -0.5 and 0.5 (indicating bias toward $\mathbf{z}_{\text{dist}}$ or $\mathbf{z}_{\text{disc}}$, respectively); 0 represents no feature bias. EvR takes values between 0 and 1 (indicating rule bias and exemplar bias, respectively).

Related formalisms and spurious correlation. This binary formulation of discriminant and distractor features has previously been studied in the context of spurious correlation (Sagawa et al., 2020b). Rather than independently varying occupancy in the four quadrants, Sagawa et al. (2020b) directly manipulate the (spurious) linear correlation between the distractor and the discriminant features (p_{maj}). In combinatorial feature spaces, a scalar spurious correlation insufficiently specifies the data distribution. The linear correlation coefficient ρ between $\mathbf{z}_{\text{disc}}$ and $\mathbf{z}_{\text{dist}}$—henceforth *spurious correlation*—can be written in terms of π_0 and π_1 via $\alpha = \frac{\pi_0 - \pi_1}{2}$ and $\beta = \frac{\pi_0 + \pi_1}{2}$ as

$$\rho(\pi_0, \pi_1) = \frac{\alpha}{\sqrt{\beta(1-\beta)}}\,. \tag{3}$$

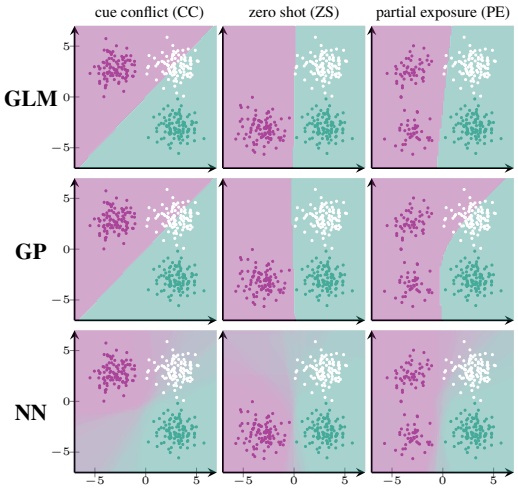
Different combinations of π_0 and π_1 give equal ρ (see the contours in Fig. (4), with markings for points along the equi-correlation contour from partial exposure ($\pi_0 = 0.5, \pi_1 = 0.0$, $\rho = 0.58$)); while nonetheless producing qualitatively different extrapolation behavior, as we demonstrate in later sections. This indicates that sensitivity to spurious correlation insufficiently specifies extrapolation behavior. We argue for a formulation like ours—based on manipulating feature *combinations*—that can tease apart distinct inductive biases at the level of what features a system finds easier to learn (FLB) as well as how to use these features to inform a decision boundary (EvR).

4. 2-D classification example

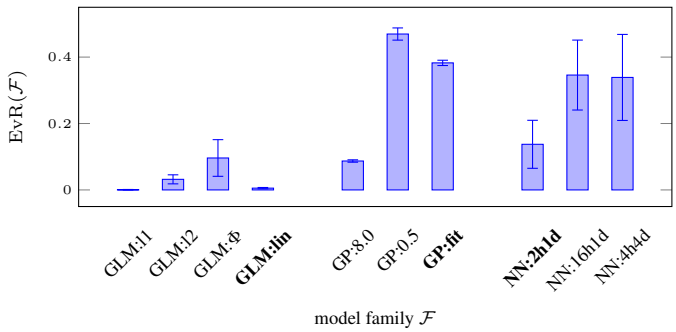
To illustrate our framework in a simple statistical learning problem and quantitatively confirm the intuitions outlined in Section (2), we consider a two-dimensional classification problem. The feature dimensions are orthogonal bases in 2D space, and we define the data generating procedure as

$$p(\mathbf{x} \mid \mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}}) = \mathcal{N}(\mu, 1.0)\,; \tag{4}$$
$$\mu = \alpha \times [2\mathbf{z}_{\text{disc}} - 1, 2\mathbf{z}_{\text{dist}} - 1]\,,$$

where, as specified in Section (3), $\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}}, \in \{0, 1\}$, $p(\mathbf{z}_{\text{disc}}, \mathbf{z}_{\text{dist}})$ is determined by the training condition. $\mathbf{z}_{\text{disc}}$ determines class labels, $\mathbf{z}_{\text{dist}}$ is a distractor, α is fixed at 3, and $N = 300$ datapoints are in each class. The group with $\mathbf{z}_{\text{dist}} = \mathbf{z}_{\text{disc}} = 1$ is assigned the test set.



(a) Decision boundaries averaged across 20 runs. Training datapoints are green or purple by label; test are white.



(b) EvR **reflects exemplar-vs-rule propensity both within and across model families.** The EvR across model families, computed across 20 runs, error bars represent 95% CIs. The GLMs are largely rule-based and show low EvR. Even within GLMs, sparsity regularization gives lower EvR. GPs are largely exemplar-based and show high EvR. Even within GPs, more 'local' GPs with lower lengthscales have higher EvR. NNs lie in-between, with larger NNs giving higher EvR.

Figure 5. **Simple 2-D classification** (Section (4)) The specific model used in (a) are bolded in (b).

4.1. Model families and nomenclature.

Neural network (NN): We train feedforward rectified linear unit (ReLU) classifiers with varying numbers of hidden layers and hidden units. We use the scikit-learn implementation with default parameters, run 20 times for confidence intervals.

Generalized linear model (GLM): Parametric models allow us to formalize the feature-sparsity that characterizes rule-based learners. Linear logistic regression is sparse by definition (it has access to only linear features). We generalize this model by expanding the feature space to include a nonlinear interaction Φ and examine L1 and L2 regularization in a GLM over this altered feature space.

Gaussian process (GP): Non-parametric kernel methods allow us to formalize exemplar-based generalization, where generalizations are made on the basis of feature-dense similarity to training data. We examine the performance of GPs with radial basis function (RBF) kernels. We fit the kernel lengthscale using gradient descent on the log marginal likelihood of the data (Rasmussen, 2003) (giving 5.2) as well as vary it (adjusting "locality" in decision boundaries); GP:8.0 denotes a GP with lengthscale value of 8.0.

We can implement explicit rule- and exemplar-based models in the synthetic setting since we know the features over which to build parametric or similarity-based models respectively, so we use it to validate our measures. In most application domains (including those in Sections (5) and (6)) feature learning is automated (Hinton & Salakhutdinov, 2006), making it difficult to specify the corresponding GLM or GP.

4.2. Comparing cue conflict, zero shot, and partial exposure

We consider one model from each class: NN with 1 hidden layer of 2 units (NN:2h1d); linear GLM (GLM:lin); RBF GP with fitted lengthscale (GP:fit). The decision boundaries learned by these models are shown in Fig. (5a). $\mathbf{z}_{\text{dist}}, \mathbf{z}_{\text{disc}}$ are equivalent by design, and permit no feature-level bias, so cue conflict is exactly at chance. This lets us focus on validating our novel protocol for measuring EvR without confounds. We generalize to cases with feature-level bias in later sections. The GLM, sparse and therefore rule-based by definition, can only learn a linear boundary. It is therefore unaffected by the distractor dimension, showing no difference in extrapolation behavior between zero shot and partial exposure (zero EvR). On the other hand, the GP is exemplar-based by definition and displays a high EvR. The NN shows an intermediate EvR, more rule-based than the purely-exemplar-based GP but not entirely rule-based like the GLM.

4.3. The influence of model properties on EvR

We first examine EvR in our control model classes (GLMs and GPs) to validate that it tracks rule- vs exemplar-based extrapolation, followed by analyses of various NNs.

Regularized GLMs: EvR **reduces with rule propensity.** A key property of rule propensity is sparsity in feature space. A linear GLM (GLM:lin) is sparse by definition, we examine a GLM on an expanded feature set so we can manipulate this sparsity. The additional feature $\Phi \propto \mathbf{z}_{\text{dist}} * \mathbf{z}_{\text{disc}}$ is the product of the observed features and normalizing by α. We compute EvR for this GLM with different regularizers

(regularization weight 1.0), shown in Fig. (**5b**).

GLM with no regularization (GLM:Φ) displays a significant EvR. L2 regularization reduces it but L1 (which directly induces feature sparsity[1]) brings it to zero (or perfectly rule-based). This demonstrates that a low EvR tracks rule propensity via feature-level sparsity.

Lengthscales in GPs: EvR increases with exemplar propensity. A sufficient condition for exemplar propensity is the locality of decision boundaries. We can directly manipulate this in a GPs with its lengthscale. We evaluate EvR in GPs with RBF kernels of different lengthscales in Fig. (**5b**). We find that the EvR is lowest with high lengthscales and grows as the lengthscale reduces, demonstrating that a high EvR tracks exemplar propensity via locality of decision boundaries.

NNs: The necessary but insufficient role of expressivity. The results from GLMs and GPs indicate that some ways to reduce expressivity (L1 regularization in GLMs and high lengthscale in RBF GPs) encourage rule propensity over exemplar propensity (thereby a lower EvR). We manipulate the most common variable in NN expressivity—its size.

We increase the width of an NN with fixed depth of 1 (Fig. (**5b**)) and find that the EvR increases. A deep NN with the same number of units, however, exhibits comparable EvR to a wide network. Deeper networks with the same number of units are more expressive than wide ones (Raghu et al., 2017), indicating that excess expressivity, while necessary, is not the sole driver of EvR.

4.4. EvR is distinct from sensitivity to spurious correlation

A crucial difference between the zero shot and the partial exposure conditions is that the partial exposure condition creates a (spurious) correlation $\rho = 0.58$ between $\mathbf{z}_{\text{dist}}$ and $\mathbf{z}_{\text{disc}}$. Is sensitivity to this spurious correlation (ρ) the sole the driver of the difference in performances between the partial exposure and zero shot conditions, *i.e.,* of the EvR? We show that this is not the case; the EvR is measuring something distinct. As described in Section (**3**), there are multiple data-settings with the same ρ. We consider training conditions specified by other π_0, π_1 that have the same ρ as the partial exposure condition (dots along the solid contour in Fig. (**4**)). We find that performance on the extrapolation quadrant after training on these new data distributions is much higher (and closer to zero shot performance) than when trained on the partial exposure condition—even though ρ is exactly the same. This indicates that performance on the partial exposure condition (normalized by zero shot performance to give the EvR) is uniquely indicative of something different from sensitivity to spurious correlations—it measures the inductive bias toward exemplar-vs-rule based extrapolation.

We can reduce ρ in different ways by increasing π_1 or by reducing π_0. We find that these are not equivalent and result in different extrapolation behaviors (*e.g.,* increasing π_1 gives more rule-based generalization than reducing π_0; see results for the 2-D classification setting in Appendix (**C.3.1**) and for the vision domain in Fig. (**7c**)). This has implications for data manipulation methods (*e.g.,* subsampling or augmentation) that manipulate this ρ to control extrapolation. This further supports that spurious correlation alone cannot explain extrapolation behavior, highlighting the importance of FLB and EvR that measure behavior under different feature *combinations* in training.

Conclusions. EvR tracks exemplar- and rule-based extrapolation, as validated on interpretable models such as GLMs and GPs. In particular, EvR decreases with reductions in expressivity mediated by regularization and lengthscale, and, in NNs, also decreases with (some kinds of) expressivity. Finally, sensitivity to spurious correlation cannot explain the EvR.

5. IMDb text classification

In this section, we demonstrate our protocol on a standard text classification task: sentiment analysis on the Internet Movie Database Movie Reviews (IMDb) dataset (Maas et al., 2011).

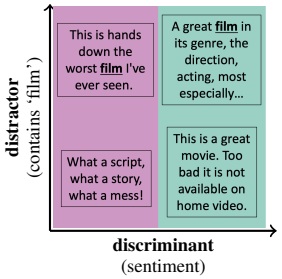


Figure 6. Example stimuli from the IMDb dataset.

Selecting features. The sentiment label ("positive" or "negative") is the discriminant $\mathbf{z}_{\text{disc}}$. We manufacture an orthogonal distractor $\mathbf{z}_{\text{dist}}$ as the presence or absence of a word that occurs in roughly 50% of the sentences in the dataset and does not occur more frequently for either positive or negative reviews. Some examples are "film" and "you": we use the word "film" (see Fig. (**6**)).

Models. We train a single layer LSTM (20 hidden units; default hyperparameters) on each condition and test on the held-out quadrant. We exclude models that do not reach 80% validation accuracy.

Feature-level bias. The distractor $\mathbf{z}_{\text{dist}}$ is easier to learn than the discriminant $\mathbf{z}_{\text{disc}}$, as reflected in the CC condition (19.7%, FLB $= -0.3$).

Exemplar bias. We see good performance in ZS (84%): Despite never having seen the word "film," the system

[1]Weight sparsity from L1-regularizer is equivalent to feature-sparsity only in special cases, including GLM.

can generalize to reviews containing it. The performance in PE drops significantly (30.1%) giving a large EvR (EvR $= 0.54$), indicating exemplar-based reasoning. As such, the exemplar-based tendency to utilize an additional unnecessary feature (*e.g.,* the presence of the word "film") hurts performance on the extrapolation quadrant.

6. CelebA image classification

We now test our protocol on a standard classification task on a large-scale image dataset, CelebFaces Attributes (CelebA) (Liu et al., 2015). Each image in this dataset is labeled with 40 binary attributes, each of which can be assigned discriminant or distractor. We examine FLB and EvR for standard models across different feature pairs, and discuss the practical implications of our findings.

Selecting features. We select feature pairs that split the data roughly evenly and thus maximizing the number of training datapoints in each quadrant. We carry out our analyses across a range of feature pairs; an example is depicted in Fig. (**7a**), and further details are in the Appendix.

Models. We train ResNets of various depths ($\{10, 18, 34\}$) and widths ($\{2, 4, 8, 16, 32, 64\}$) on 6 different choices for feature pairs, with standard hyperparameters. We limit our analyses to networks that achieve at least 75% validation accuracy (on held-out samples from its own training distribution) to ensure that, despite differences in data variability across training conditions, all models learn a meaningful decision boundary.

Feature-level bias. There is a wide range of FLB across feature pairs; *e.g.,* "male" is easier to learn than "high cheekbones" giving high FLB, and "mouth open" and "wearing lipstick" are equally difficult and give FLB of close to 0. FLB values for each feature pair were consistent across ResNet widths and depths.

Exemplar-rule bias. We observe good ZS performance: the models can generalize to new feature values outside the training support. We see a wide range of EvR across feature pairs, Fig. (**7b**). Across all feature pairs, the EvR is non-negative: generalization in the PE condition is always worse (or not significantly better) than in the ZS condition. Further, we see a linear correlation between EvR and FLB in logit space across feature pairs. EvR therefore depends on how easy or hard the features are to learn. The key, however, is that this regression of the EvR onto FLB has a positive intercept: there is a positive EvR even for feature pairs with no FLB. That is, we see lower performance in PE compared to ZS (a nonzero EvR, exemplar propensity) even when FLB is controlled for.

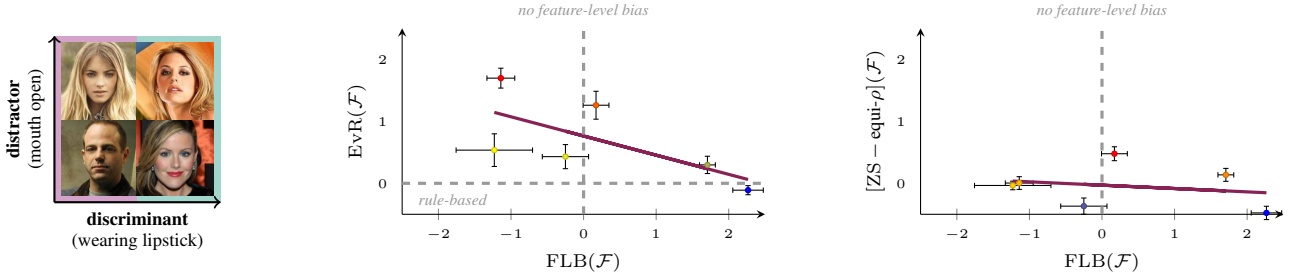
We find no differences in EvR across ResNet widths and depths: Fig. (**7b**) plots EvR and FLB averaged over ResNet sizes (model-specific results in Appendix). One explanation is that the features in CelebA are complex; to learn these, we need reasonably high model expressivity, and differences in parameter count do not further modulate EvR. This is consistent with findings in Section (**4**) where expressivity is necessary but not sufficient for increases in EvR: we see a jump in EvR going from NN:2h1d to NN:16h1d, but no further change going to the even more expressive NN:4h4d.

Controlling spurious correlation. We replicate the findings in Section (**4**): the EvR cannot be explained by sensitivity to spurious correlation ρ. This is demonstrated in Fig. (**7c**), where we substitute performance in the PE condition with performance in a different data condition ($\pi_0 = 0.825, \pi_0 = 0.25$) with the same $\rho = 0.58$ as in the PE condition. We find none of the effects discussed above, indicating that the PE condition is measuring something unique—exemplar-vs-rule propensity—which is not accounted for by sensitivity to spurious correlation. Further, EvR does not increase with model expressivity, unlike sensitivity to spurious correlation (Sagawa et al., 2020b).

Practical implications of the EvR**.** The nonzero EvR (*i.e.,* exemplar bias) reveals that models are better at extrapolating zero-shot to a new feature value than when they have partial exposure to that feature value *even though the additional data need not change the learned decision boundary*. In particular, the training examples added in PE can be classified with the decision function from ZS without incurring additional training loss. A rule-based system recognizes this and bases its generalization on the minimal features that support the category boundary. However, an exemplar-based model changes its decision boundary in response to this additional data.

PE-approximating data distributions ($\pi_0 \approx 0.5, \pi_1 \approx 0.0$) occur naturally. For example, as (Sagawa et al., 2020b) observe, "blond" "male"s are under-represented in CelebA. Consistent with the rest of our results, we find better classification for the extrapolation quadrant (blond males) if we discard data from an adjacent quadrant (blond non-males, or non-blond males) simulating the zero-shot condition, as opposed to the PE condition if such data is included: ResNet10, width 2, gives ZS $= 75.12 \pm 3.09\%$; PE $= 60.22 \pm 7.27\%$ for $\mathbf{z}_{\text{disc}}$ ="male" (discard blond non-males to get ZS) and ZS $= 68.16 \pm 3.34\%$; PE $= 49.78 \pm 3.76\%$ for $\mathbf{z}_{\text{disc}}$ ="blond" (discard non-blond males to get ZS).

These results demonstrate the practical impact of understanding the exemplar-vs-rule bias in a model: an exemplar biased model (like the ResNet here) generalizes poorly in combinatorial settings, and can be made to generalize better by discarding an entire quadrant of data. Previous sub-sampling approaches (Sagawa et al., 2020b; Haixiang et al., 2017) do not manipulate feature combinations and only manipulate spurious correlations. The aforementioned



(a) Example CelebA stimuli; we test 6 discriminant-distractor pairs (with sufficient coverage of quadrants in Fig. (3)) of 6 features.

(b) Main effect, logit scale: Exemplar-vs-rule propensity (EvR) trends linear with non-zero intercept (fit in red) in feature-level bias (FLB).

(c) Control, logit scale: Performance deviation from ZS under equi-correlation interpolation is close to zero; linear fit (red) has intercept at zero.

Figure 7. **CelebA results.** Stimuli and results on various feature pairings from the CelebA domain (Section (6)). Error bars represent 95% confidence intervals across ResNets of various sizes. See figure sub-captions and main text for details.

analyses (Fig. (4)) and results (Fig. (7c)) demonstrate that this underspecifies extrapolation behavior.

7. Related work and future directions

Model design for systematic generalization. Rule-based generalization permits systematic extrapolation in combinatorial domains. This systematicity has been found lacking in neural networks (Lake & Baroni, 2018b; Barrett et al., 2018), leading to renewed interest in hybrid symbolic–connectionist methods (*e.g.,* Garnelo & Shanahan, 2019). However, works proposing new methods usually do not examine how feature co-occurrences modulate the systematicity of extrapolation. Using our protocol to examine exemplar- vs. rule-based generalization in these models is a promising future direction.

Learning causal features. Rule-based generalization, is equivalent to learning causal features under the assumption that the causal model is the simplest model that explains the data. Recent work has investigated data settings that separate causal features from spurious ones (*e.g.,* Arjovsky et al., 2019).We showed that a model with exemplar propensity makes more rule-based extrapolations for certain training feature combinations (*i.e.,* zero shot vs. partial exposure). Investigating how feature coverage impacts causal generalization is a fruitful future direction.

Similarity-based generalization and kernels. We use similarity-based kernels (*e.g.,* radial basis function (RBF)) to exemplify exemplar-based extrapolation. Recent work has interpreted neural networks as kernel regression (Jacot et al., 2018). Using a kernel framing to formalize the causes of exemplar bias is an exciting future direction.

Data augmentation. The EvR measure allows us to demonstrate that increased data variation in the form of feature coverage worsens systematic generalization. The negative effect of data variation on generalization has been documented for adversarial augmentations (Raghunathan et al., 2020). We show that this can persist even when augmentation is not adversarial, rendering it generally relevant for the design of data augmentations.

8. Conclusions

Taking inspiration from—and going beyond—psychological studies, we design a behavioral protocol to distinguish the effects of two inductive biases (feature-level bias and exemplar bias) that is easily applicable to any classification domain. This follows in a promising line of recent work that analyses and interprets deep learning systems based on their external behavior (Ritter et al., 2017b; Dasgupta et al., 2019). It complements other approaches that follow in the neuroscience tradition of analyzing internal representations (Zeiler & Fergus, 2014; Karpathy et al., 2015) or make approximations of these internal workings to support theoretical results (Jacot et al., 2018; Allen-Zhu et al., 2019). The behavioral approach has the advantage that it makes no assumptions about the model, allowing comparisons across systems that differ in design.

Both rule- and exemplar-based extrapolation are valuable depending on domain, underscoring the importance of diagnosing feature-level bias and exemplar bias. Moreover, studying this trade-off allows us to demonstrate an important phenomenon: We find that more feature coverage (as in partial exposure compared to zero shot) hurts generalization for exemplar-based models. This has implications for methods that manipulate data distributions to improve performance (*e.g.,* data subsampling (Haixiang et al., 2017), data augmentation (Perez & Wang, 2017), and contrastive learning (Chen et al., 2020)). Since an exemplar-based model tends to acquire spurious associations, our measures have the potential to be useful as diagnostics in application settings where the goal is to control model behavior on non-representative factors ((*e.g.,* Mitchell et al., 2019)).

A limitation of the present work is that we do not provide a conclusive answer as to what properties of a model family influence both feature-level bias and exemplar bias. A broader study of these factors and theoretical work formalizing this effect are exciting avenues for future work.

Acknowledgements and Funding Sources

We thank our anonymous reviewers for their feedback. This work was supported by ONR Grant #N00014-18-1-2873 and the DARPA L2M program.

References

Allen, S. W. and Brooks, L. R. Specializing the operation of an explicit rule. *Journal of experimental psychology: General*, 120(1):3, 1991.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

Andreas, J. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

Andreas, J., Rohrbach, M., Darrell, T., and Klein, D. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48, 2016. URL https://arxiv.org/abs/1511.02799.

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

Ashby, F. G. and Townsend, J. T. Varieties of perceptual independence. *Psychological Review*, 93(2):154, 1986.

Barrett, D., Hill, F., Santoro, A., Morcos, A., and Lillicrap, T. Measuring abstract reasoning in neural networks. In *International Conference on Machine Learning*, pp. 511–520, 2018. URL https://arxiv.org/abs/1807.04225.

Beery, S., Van Horn, G., and Perona, P. Recognition in terra incognita. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018. URL https://arxiv.org/abs/1807.04975.

Brown, G., Bun, M., Feldman, V., Smith, A., and Talwar, K. When is memorization of irrelevant training data necessary for high-accuracy learning? *arXiv preprint arXiv:2012.06421*, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020. URL https://arxiv.org/abs/2011.03395.

Dasgupta, I., Guo, D., Gershman, S. J., and Goodman, N. D. Analyzing machine-learned representations: A natural language case study. *arXiv preprint arXiv:1909.05885*, 2019.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Feldman, V. and Zhang, C. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33, 2020.

Garnelo, M. and Shanahan, M. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2018.12.010. URL https://www.sciencedirect.com/science/article/pii/S2352154618301943. Artificial Intelligence.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

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

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239, 2017.

Hermann, K. L., Chen, T., and Kornblith, S. The origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071*, 2019.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *science*, 313 (5786):504–507, 2006.

Hudson Kam, C. L. and Newport, E. L. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language learning and development*, 1(2):151–195, 2005.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2901–2910, 2017.

Karpathy, A., Johnson, J., and Fei-Fei, L. Visualizing and understanding recurrent networks. *Proceedings of the 3rd International Conference on Learning Representations, Workshop Track*, 2015.

Lake, B. and Baroni, M. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International Conference on Machine Learning*, pp. 2873–2882. PMLR, 2018a.

Lake, B. M. and Baroni, M. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *Proceedings of the 6th International Conference on Learning Representations, Workshop Track*, 2018b.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2018.

Landau, B., Smith, L. B., and Jones, S. S. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.

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

Maas, A., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Marcus, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 220–229, 2019.

Mitchell, T. M. *The need for biases in learning generalizations*. Department of Computer Science, Laboratory for Computer Science Research, 1980.

Nosofsky, R. M., Clark, S. E., and Shin, H. J. Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, memory, and cognition*, 15(2):282, 1989.

Perez, L. and Wang, J. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

Pothos, E. M. The rules versus similarity distinction. *Behavioral and brain sciences*, 28(1):1, 2005.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *international conference on machine learning*, pp. 2847–2854. PMLR, 2017.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J., and Liang, P. Understanding and mitigating the tradeoff between robustness and accuracy. *arXiv preprint arXiv:2002.10716*, 2020.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Rips, L. J. Similarity, typicality, and categorization. 1989.

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. Cognitive psychology for deep neural networks: A shape bias case study. *International Conference on Machine Learning*, pp. 2940–2949, 2017a.

Ritter, S., Barrett, D. G., Santoro, A., and Botvinick, M. M. Cognitive psychology for deep neural networks: A shape bias case study. In *International conference on machine learning*, pp. 2940–2949. PMLR, 2017b. URL https://arxiv.org/abs/1706.08606.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020a. URL https://arxiv.org/abs/1911.08731.

Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pp. 8346–8356. PMLR, 2020b.

Shepard, R. N. and Chang, J.-J. Stimulus generalization in the learning of classifications. *Journal of Experimental Psychology*, 65(1):94, 1963.

Smith, E. E. and Sloman, S. A. Similarity-versus rule-based categorization. *Memory & Cognition*, 22(4):377–386, 1994.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. *European conference on computer vision*, pp. 818–833, 2014.

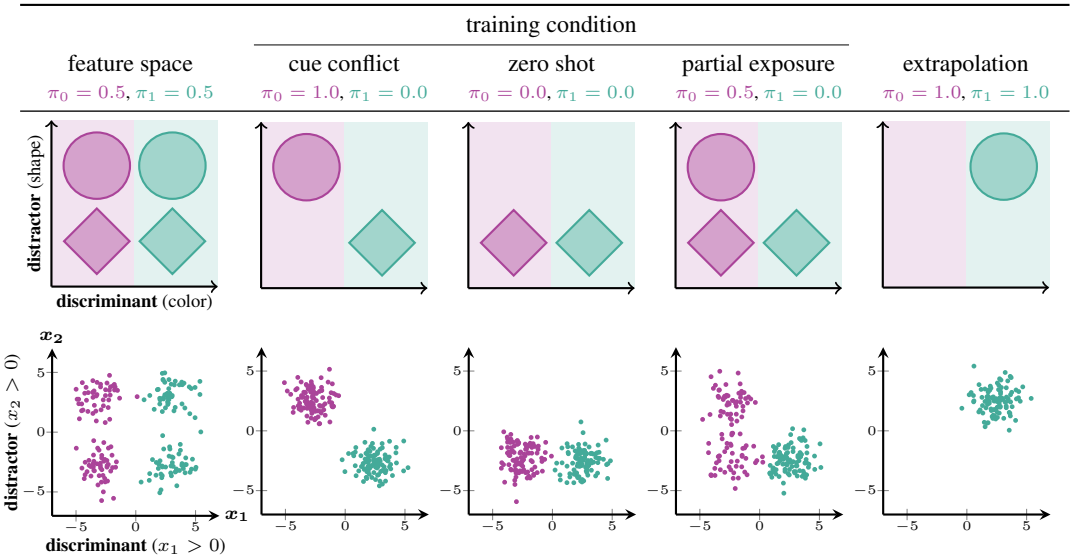


Figure 8. We expand on Fig. (3) from the main text by including a realization of the abstract training conditions in the simple 2D points-in-a-plane setting. (**Top**) **Formalizing the illustrative experiment:** The experiment from Fig. (2) expressed in terms of the formalism in Section (3) with $\mathbf{z}_{\text{dist}}$ = color and $\mathbf{z}_{\text{disc}}$ = shape. Background colors indicate true category boundary. (**Bottom**) The conditions realized via a binarization of continuous feature values. Here, the discriminant is binarized as $x_1 > 0$ and the distractor as $x_2 > 0$; this setting is further investigated in Section (4). Color here depicts the label but is not part of the input.

A. Additional formalizations

A.1. Generalizing the framework from two binary attributes to many categorical attributes

In the most general terms, we consider a setting in which each observation $\mathbf{x} \in \mathcal{X}$ is underlied by n categorical variables $z_1, \ldots, z_n \in \{0, \ldots, C\}$ with $C \in \mathbb{Z}_+$, henceforth *attributes* whose concatenation $\mathbf{z} = (z_1, \ldots, z_n)$ determines the observable input $\mathbf{x}$ via some mapping $g : \mathbb{Z}_{0+}^n \to \mathcal{X}$. We consider the binary classification task of fitting a model $\hat{f} : \mathcal{X} \to \{0, 1\}$ from a given model family $\mathcal{F}$ to predict a binary label for each input. A subset of the attributes in $\mathbf{z}$, without loss of generality $(z_0, \ldots, z_i)$, is taken to define the decision boundary, while the remaining attributes, $z_{i+1}, \ldots, z_n$, are assumed to not be independently predictive of the true classification $y \in \{0, 1\}$. We therefore denote the *discriminant*, $\mathbf{z}_{\text{disc}} = (z_0, \ldots, z_i)$, and the *distractor* $\mathbf{z}_{\text{dist}} = (z_{i+1}, \ldots, z_n)$. For simplicity, we assume that the attributes are binary (*i.e.,* $C = 2$ and $z_i \in \{0, 1\}, \forall i$), and that the discriminant attributes must be jointly active for the classification to change from the null class $y = 0$ (*i.e.,* $y = 1 \iff \mathbf{z}_{\text{disc}} = \mathbf{1}$); the latter simplification allows us to redefine $\mathbf{z}_{\text{disc}} = z_0 \wedge \cdots \wedge z_i$ and $\mathbf{z}_{\text{dist}} = z_{i+1} \wedge \cdots \wedge z_n$, which is equivalent to the earlier discussion of the illustrative two-attribute case.

A.2. Training conditions expressed in terms of the joint distribution

We express the training conditions displayed in Fig. (3) and realized in Figure 6 in terms of the joint distribution instead of the parameters π_0, π_1.

1. The cue-conflict condition the upper left and lower right quadrants in Figure 6 and defines the distribution of attributes as

$$\begin{array}{l|l} p_{\text{cc}}(\mathbf{z}_{\text{disc}} = 0, \mathbf{z}_{\text{dist}} = 1) = 0.5 & p_{\text{cc}}(\mathbf{z}_{\text{disc}} = 1, \mathbf{z}_{\text{dist}} = 1) = 0 \\ \hline p_{\text{cc}}(\mathbf{z}_{\text{disc}} = 0, \mathbf{z}_{\text{dist}} = 0) = 0 & p_{\text{cc}}(\mathbf{z}_{\text{disc}} = 1, \mathbf{z}_{\text{dist}} = 0) = 0.5\,. \end{array}$$

2. The zero-shot condition populates the bottom left and right quadrants in Figure 6 and defines the distribution of attributes as

$$\begin{array}{l|l} p_{\text{zs}}(\mathbf{z}_{\text{disc}} = 0, \mathbf{z}_{\text{dist}} = 1) = 0 & p_{\text{zs}}(\mathbf{z}_{\text{disc}} = 1, \mathbf{z}_{\text{dist}} = 1) = 0 \\ \hline p_{\text{zs}}(\mathbf{z}_{\text{disc}} = 0, \mathbf{z}_{\text{dist}} = 0) = 0.5 & p_{\text{zs}}(\mathbf{z}_{\text{disc}} = 1, \mathbf{z}_{\text{dist}} = 0) = 0.5\,. \end{array}$$

3. The partial-exposure condition populates all quadrants but the upper right in Figure 6 and defines the distribution of attributes as

$$\begin{array}{l|l} p_{\text{pe}}(\mathbf{z}_{\text{disc}}=0,\mathbf{z}_{\text{dist}}=1)=0.25 & p_{\text{pe}}(\mathbf{z}_{\text{disc}}=1,\mathbf{z}_{\text{dist}}=1)=0 \\ \hline p_{\text{pe}}(\mathbf{z}_{\text{disc}}=0,\mathbf{z}_{\text{dist}}=0)=0.25 & p_{\text{pe}}(\mathbf{z}_{\text{disc}}=1,\mathbf{z}_{\text{dist}}=0)=0.5\,. \end{array}$$

B. More CelebA results

We include model-specific results, split by ResNet depth and width, in Fig. (9). We find no systematic relationship between EvR and depth or width.

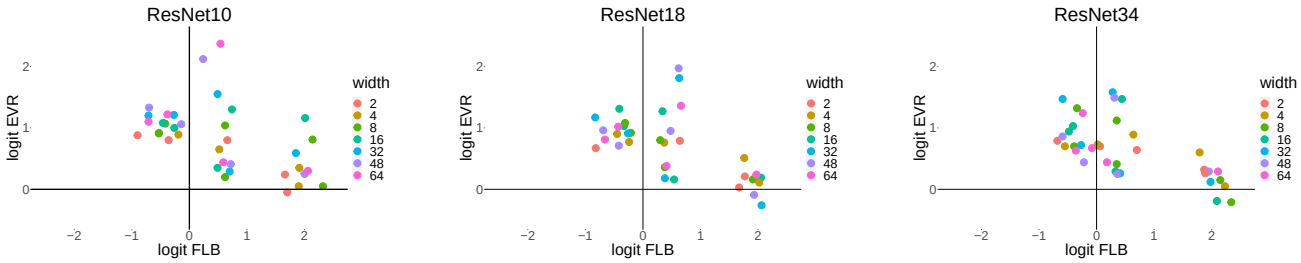


Figure 9. CelebA EvR and FLB across feature pairs, averaged across 30 runs, split by depth and width of ResNet.

C. Spurious correlation underdetermines feature distributions

The partial-exposure condition ($\pi_0 = 0.5, \pi_1 = 0.0$) in Section (2) results in a spurious correlation between the discriminant $\mathbf{z}_{\text{disc}}$ and the distractor $\mathbf{z}_{\text{dist}}$ ($\rho = 0.58$). To examine behavior in a wider range of data settings, we vary π_0 and π_1 as described in Section (3), thereby also changing the degree of spurious correlation.

I. Interpolation towards zero shot. We interpolate π_0 from 0.5 towards 0.0, keeping $\pi_1 = 0.0$. This moves us closer to $\pi_0 = \pi_1 = 0.0$, where we have no exposure to $\mathbf{z}_{\text{disc}} = 1$ in training. Intuitively, we are reducing the exposure to the new distractor feature value from the partial-exposure condition.

II. Interpolation to full exposure. We interpolate π_1 from 0.0 towards 0.5, keeping $\pi_0 = 0.5$. This moves us closer to $\pi_0 = \pi_1 = 0.5$, where we have equal exposure to all quadrants in training. Here, rather than reducing the exposure to the new distractor feature value, we are equalizing the exposure to it across the discriminant dimension.

III. Interpolation with matched correlation. We report results on this in Sections (4) and (6). As also depicted in Fig. (4), we generate training conditions by changing π_0 and π_1 such that we follow a ρ-contour away from the partial-exposure condition ($\pi_0 = 0.5, \pi_1 = 0.0, \rho = 0.58$): solid contour in Fig. (4). We also match the spurious correlation across the two interpolations in Appendix (C)A and B: Fig. (11) shows these additional ρ-contours as dashed lines.

These different interpolations are depicted in Fig. (11a) with different shape/colors.

C.1. Generating interpolation points

We generate points along all three interpolation lines: from partial exposure towards zero shot ((C)I); from partial exposure towards full exposure ((C)II); and the equi-correlation line originating from partial exposure ((C)III). The interpolating points along each line are selected to balance spurious correlation and feature exposure. In particular, we follow the following procedure:

1. We choose a point that interpolates towards full exposure. We do this by choosing a value of π_1 between 0.0 and 0.5, π^{FE}. This gives a data setting, along with a corresponding spurious correlation, ρ, computed via Eq. (3):

$$\pi_0 = 0.5\,; \qquad \pi_1 = \pi^{\text{FE}}\,; \qquad \rho = \rho\left(0.5, \pi^{\text{FE}}\right)\,.$$

2. We generate a corresponding point that interpolates towards zero shot. Given the data setting above, we set $\pi_1 = 0.0$ and compute the π_0 to produce the same ρ as the full-exposure interpolations in Step 1. This gives the data setting:

$$\pi_0 = \pi^{\text{ZS}}\left(\pi^{\text{FE}}\right)\,; \qquad \pi_1 = 0.0\,; \qquad \rho = \rho\left(\pi^{\text{ZS}}\left(\pi^{\text{FE}}\right), 0.0\right) = \rho\left(0.5, \pi^{\text{FE}}\right)\,.$$

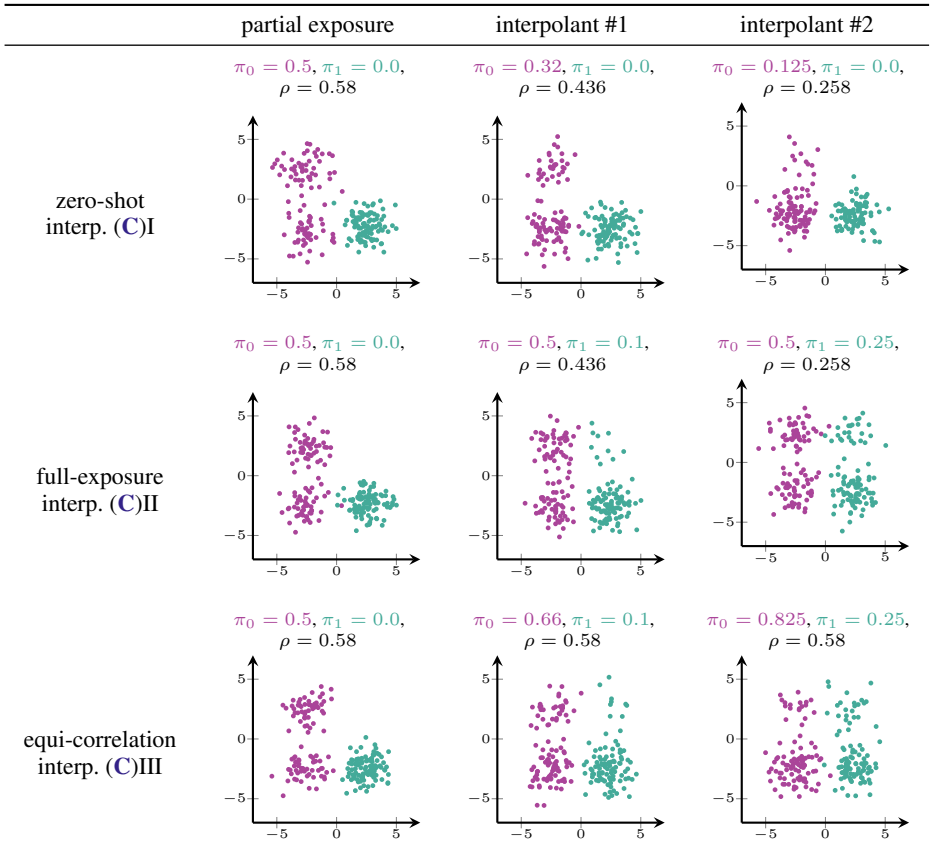


Figure 10. We visualize several of the interpolants used for the interpolation analyses.

3. Finally, we also derive the equi-correlation interpolation from the full-exposure interpolation as follows. We retain π_1 from the full-exposure condition, but recompute the π_0 such that the correlation ρ matches the spurious correlation of the pure
glspec ($\rho = 0.58$). This gives an additional data setting:

$$\pi_0 = \pi^{\text{EQ}}\left(\pi^{\text{FE}}\right) \; ; \qquad \pi_1 = \pi^{\text{FE}}; \qquad \rho = \rho\left(0.5, 0.0\right) = 0.58 \; .$$

Note that, despite there being three different interpolation lines, the specific interpolants we use are constrained along a single degree of freedom—choosing π^{FE} (Step 1). The data settings for zero shot (Step 2) and equi-correlation (Step 3) are derived from this value.

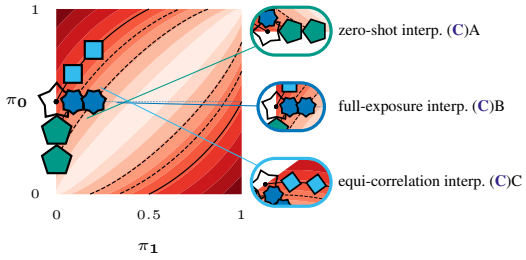
C.2. Specific interpolation values used

For all data settings, we generate points along the interpolation lines using the procedure in Appendix (**C.1**).

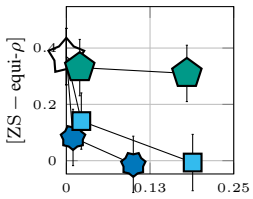
For the simple 2D classification setting, we examine two interpolants. In this simple domain, we keep the interpolation distances small, since we expect changes in extrapolation behavior even from small changes.

	interpolant 1			interpolant 2		
	π_0	π_1	ρ	π_0	π_1	ρ
interpolation to zero shot ((C)I)	0.481	0.0	0.563	0.32	0.0	0.436
interpolation to full exposure ((C)II)	0.5	0.01	0.563	0.5	0.1	0.436
equi-correlation interpolation ((C)III)	0.519	0.01	0.58	0.661	0.1	0.58

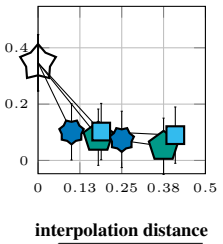
For CelebA, we increase the interpolation distance to reflect the wider range of natural data distributions among feature



(a) A heatmap of the spurious correlation (Eq. (3)), showing different interpolations. The partial-exposure condition is identified with a star outline; points along the three interpolation line are identified with filled shapes.



(b) Interpolations for NN:16h1d on 2D classification of points-in-a-plane.



(c) Interpolations, ResNet-18-8, CelebA ("wearing lipstick", "mouth open")

Figure 11. Interpolations away from the PE: changes in extrapolation behavior under data distribution with the same spurious correlation as in PE, as well as different ways to change spurious correlation.

pairs. The data these interpolation values generate is visualized as the equivalent points-in-a-plane setting in Figure 7.

	interpolant 1			interpolant 2		
	π_0	π_1	ρ	π_0	π_1	ρ
interpolation to zero shot ((C)I)	0.32	0.0	0.436	0.125	0.0	0.258
interpolation to full exposure ((C)II)	0.5	0.1	0.436	0.5	0.25	0.258
equi-correlation interpolation ((C)III)	0.66	0.1	0.58	0.825	0.25	0.58

C.3. Interpolation analyses

C.3.1. In the 2-D classification example

In the simple setting from Section (4), we vary π_0, π_1 for an NN model (NN:16h1d, the NN with lowest EvR level overall). Results are in Fig. (11b) and discussed below.

EvR $\neq$ **sensitivity to spurious correlation.** As also discussed in the main text, along the equi-correlation interpolation line, the "effective EvR" drops drastically (*i.e.,* the learner generalizes in more rule-based manner) despite no change in spurious correlation.

Implications for controlling extrapolation. Despite both having the same ρ, interpolating towards full-exposure increases the EvR more than towards zero-shot. This further supports that spurious correlation cannot fully characterize extrapolation behavior. This shows that different *ways* to reduce ρ have different effects on extrapolation, and has important implications for data manipulation methods (*e.g.,* subsampling or augmentation) that aim to directly control this ρ.

C.3.2. In CelebA

We see the same effects as in the linear setting: as also discussed in the main text, we see a much smaller gap to the ZS condition despite no change in spurious correlation. We don't find clear effects distinguishing different ways to reduce spurious correlation (interpolation to zero shot ((C)I) and interpolation to full exposure ((C)II)).