# Frequency based pruning improves latent identification and generalization

**Ashish Malik** [1] **Alan Fern** [1]

## Abstract

In this work, we propose Frequency-Based Pruning (FBP), a novel method for mitigating spurious correlations and improving latent variable identification in machine learning models, particularly under weak supervision. Spurious correlations—irrelevant features coincidentally predictive of outcomes—degrade model generalization and robustness, especially in biased or noisy data scenarios. To address this, we introduce Activation Variability-Based Pruning (AVP), a strategy that selectively removes parameters with low activation variability, hypothesizing these to align with spurious signals.

Building on existing results that latent variables can be identified up to an affine transformation under diverse perturbations, we demonstrate that pruning strategies enhancing activation variability reduce spurious correlations and improve robustness. Our method unifies spurious effects arising from structured data biases and random label noise into a single framework, showing its effectiveness in isolating robust features across diverse weakly supervised settings. Empirical results validate our approach on both synthetic and real-world datasets. Compared to baselines, our method consistently achieves improved robustness and generalization. This work highlights the importance of activation variability as a signal for pruning, providing a principled and effective strategy for enhancing model performance in weakly supervised environments.

## 1. Introduction

Machine learning models often suffer from spurious correlations—irrelevant features that are predictive on training data but fail to generalize under distribution shifts. These correlations undermine model robustness and generalization, particularly in weakly supervised settings where spurious signals can hinder or even dominate the learning process. For example, models trained on natural datasets often exploit shortcuts like backgrounds or textures, instead of learning robust, causal features. Such failures hinder performance on out-of-distribution (OOD) data, a critical challenge for deploying machine learning models in real-world applications.

The identification of latent variables which are the true, underlying factors generating the observed data is quintessential to building robust models. Prior works have shown that true latent representations can be recovered up to affine transformations under specific perturbations (Ahuja et al., 2022). However, models often struggle to isolate these core representations from spurious features in weakly supervised settings. Many recent advances have highlighted the importance of filtering out non-informative signals. Despite these advances, existing methods lack principled strategies to systematically reduce reliance on spurious correlations while retaining robust latent representations.

Pruning techniques, which are traditionally used for model compression, have emerged as a promising strategy to improve generalization and robustness (Chen et al., 2022; Hoefler et al., 2021). Sparse networks often exhibit better generalization by acting as implicit regularizers. However, sparsity alone does not guarantee robustness; in fact, improper pruning can worsen overfitting, leading to phenomena such as sparse double descent (He et al., 2022). Existing pruning methods also fail to explicitly target spurious correlations, which remain embedded in latent representations.

In this work, we propose a novel approach to improve latent variable identification and generalization in weakly supervised settings via pruning. Unlike conventional pruning, we leverages activation variability as a signal to identify and remove parameters corresponding to spurious correlations. We hypothesize that parameters with low activation variability contribute to spurious features, and their removal enhances the model's reliance on robust, informative signals. Our method builds upon recent insights that sparsity can improve generalization and that latent variables can be identified up to affine transformations under controlled perturbations (Ahuja et al., 2022).

[1]College of Engineering, Oregon State University, USA. Correspondence to: Ashish Malik <malikas@oregonstate.edu>.

*Proceedings of the $41^{st}$ International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

This work makes the following contributions:

- **Theoretical Foundations**: We extend prior results on latent variable identification and show that pruning strategies enhancing activation variability reduce reliance on spurious features.
- **Unified Framework for Weak Supervision**: We unify spurious correlations caused by structured biases and random label noise under a single framework, demonstrating that frequency-based pruning mitigates both.
- **Empirical Validation**: Through experiments on synthetic and real-world datasets, we show that our approach consistently improves robustness and generalization compared to state-of-the-art baselines.

Our work highlights the critical role of activation variability in isolating robust features and mitigating spurious correlations. By bridging theoretical insights with practical pruning strategies, we provide a principled and effective method for enhancing model performance in weakly supervised and out-of-distribution settings.

## 2. Related works

**Latent Variable Identification and Robustness:** Identifying latent variables in the presence of spurious correlations is critical for model robustness, particularly in weakly supervised settings. Recent studies demonstrate that introducing controlled perturbations improves the separation of relevant and spurious features. (Ahuja et al., 2022) highlight this in weakly supervised representation learning, using sparse perturbations to enhance robust feature isolation. Similarly, counterfactual invariance frameworks for classification emphasize the need for models that maintain stable predictions across shifts by ignoring irrelevant features (Victor et al., 2021). Our approach leverage these insights by reducing the impact of spurious signals via selective pruning, which refines latent identification and improving generalization.

**Pruning and Sparsity for Improved Generalization:** Pruning has shown promise for enhancing both efficiency and robustness by focusing models on relevant features. For instance, the "Sparse Double Descent" phenomenon explores how high sparsity levels mitigate over-fitting and support generalization (He et al., 2022). Further, (Chen et al., 2022) demonstrates that structured pruning not only narrows the generalization gap in adversarial settings but also improves standard performance. These findings align with our use of frequency-based pruning, which acts as an implicit regularizer that enhances robustness.

**Affine Transformations for Latent Representations:** Prior works have demonstrated that ERM-based models capture the core features even when heavily influenced by spurious signals, allowing for minor-retraining to improve generalization (Izmailov et al., 2022). (Tripuraneni et al., 2021) adds that over-parameterized models tend to exhibit a stable affine relationship between in-distribution and out-of-distribution generalization. This work extends these highlights to show that pruning can be used to preserve the core affine transformations in the latent space while filtering out spurious features to maintain robust performance without over-parameterization.

Other works on generalization theory, such as by (Nagarajan & Kolter, 2019) and (Ilyas et al., 2019) reveals that standard measures often overlook the role of non-robust features that undermine performance in high-dimensional and adversarial settings. This underscores the value of this work that selectively reduce model reliance on such non-informative and often disruptive features. Through selective feature pruning, we aim to isolate core latent structures, advancing model robustness and stability in environments prone to spurious correlations.

## 3. Feature robustness with spurious data

Our formalism follows (Ahuja et al., 2022) with some important changes. Consider two classes of variables - a) *Observed* variables $X \in \mathcal{X} \subseteq \mathbb{R}^n$ and b) latent variables $Z \in \mathcal{Z} \subseteq \mathbb{R}^d$, where $\mathcal{X}$ and $\mathcal{Z}$ are the set of all possible observed and latent variables. The latent variables $Z$ are sampled from some unknown distribution $\mathbb{P}_Z$. Let $g : \mathbb{R}^d \to \mathbb{R}^n$ be a an injective and analytic function that generates $X$ using $Z$. Let $\mathcal{I} = \{I_1, I_2, ...., I_m\}$ be a set of possible perturbations applicable on $Z$ and the corresponding perturbation vectors $\Delta = \{\delta_1, \delta_2, ..., \delta_m \mid \delta_i \in \mathbb{R}^d\}$. Finally, let $h : \mathbb{R}^d \to \mathbb{R}^e$ be a bijective mapping that maps observed vectors $X$ to a set of labels $Y$. Formally, the data generation process can be described as follows:

**Assumption 3.1.** DGP: The data generation process is:

$$
\begin{aligned}
&z \sim \mathbb{P}_Z, \\
&\tilde{z}_i \leftarrow z + \delta_i, \quad \forall \delta_i \in \Delta \\
&\tilde{x}_i \leftarrow g(\tilde{z}_i), \quad \forall \delta_i \in \Delta \\
&x \leftarrow g(z), \\
&y \leftarrow h(x)
\end{aligned}
$$

Here, $x, \tilde{x}, y$ and $z$ are realization of $\mathcal{X}, X, Y$ and $Z$ respectively. It is important to note that the observed vectors are generated using perturbed latents, however, the corresponding labels are generated using the same latents but unperturbed.

Let $D \in \{\tilde{x} \leftarrow g(z+\delta_i), y \leftarrow h \circ g(z) | \forall \delta_i \in \Delta, \forall z \sim \mathbb{P}_z\}$ be all the variable-label pairs that can be created using DGP described in 3.1 and let $\tilde{D} = \{(\tilde{x}, y)\} \subseteq D$ be the set of observed variable label pairs.

**Learning problem:** Assume that the learning problem is to learn a function $Q : \mathbb{R}^n \to \mathbb{R}^e$ that correctly maps $\tilde{X} \in \{g(z+\delta_i)|\forall\delta_i \in \Delta, z \sim \mathbb{P}_z\}$ to their corresponding $Y \in \{h \circ (z)|z \sim \mathbb{P}_Z\}$. This function is learned using $\{\tilde{x}, y\} \in \tilde{D}$. Let's assume that $Q$ consist of two components: a feature extractor $f : \mathbb{R}^n \to \mathbb{R}^d$ and classifier $c : \mathbb{R}^d \to \mathbb{R}^e$. Let $f$ be parameterized using parameters $\theta_f$ and $c$ be parameterized using $\theta_c$. The learning objective is to minimize the cross entropy loss function:

$$\mathcal{L}(\theta_f, \theta_c) = \mathbb{E}_{\tilde{x},y\sim\tilde{D}}\left[-\log\left(\frac{\exp\Big(c\big(f(\tilde{x})\big)\Big)_y}{\sum_{j=1}^{e}\exp\Big(c\big(f(\tilde{x})\big)\Big)_j}\right)\right] \tag{1}$$

**Assumption 3.2.** Oracle classifier: Let $c^* : \mathbb{R}^d \to \mathbb{R}^e$ be the oracle classifier that can correctly map all $z \in Z$ to the corresponding label $y \in Y$.

Assuming access to the oracle classifier $c^*$ for $Q$, the output of the feature extractor $f$ at any point during learning is:

$$\begin{aligned} f(\tilde{x_i}) &= g^{-1}(\tilde{x}_i) + \epsilon_i, \quad \forall(\tilde{x}_i, y_i) \in D \\ &= c^{*-1}(y_i) + \epsilon_i, \quad \forall(\tilde{x}_i, y_i) \in D \\ &= f(x_i) + \upsilon_i, \quad \forall(\tilde{x}_i, y_i) \in D \end{aligned} \tag{2}$$

Here, $\epsilon_i, \upsilon_i \in \mathbb{R}^d$. $\upsilon_i$ is the excess noise generated by the feature extractor $f$. Since $h$ is bijective, from assumption 3.2, it follows that $c^{*-1}$ exists. Let $\Upsilon = \{\upsilon_i\}_{\forall i}$, so, the best possible feature extractor $f^*$, minimizes the following error:

$$f^* = \min_{\theta_f} \mathbb{E}_{[D,\Upsilon]}\left[\left(f(\tilde{x}_i) - f(x_i) - \upsilon_i\right)^2\right] \tag{3}$$

Note that the ideal feature extractor $f^*$ may not be learnable due to the training dynamics, expressivity limitations, or $\mathbb{P}((x,y) \sim \tilde{D}) \neq \mathbb{P}((x,y) \sim D)$, etc. Let $\hat{f}$ represent the best learnable feature extractor. Let the set of latents output by the $\hat{f}$ be $\hat{Z} = \{\hat{z} \leftarrow \hat{f}(x)|\forall x \in X\}$. The predicted latents $\hat{z} \in \hat{Z}$ are related to the true latents $z \in Z$ as follows:

$$\begin{aligned} \hat{z} &= \hat{f} \circ g(z), \forall z \in Z \\ &= a(z), \forall z \in Z \end{aligned} \tag{4}$$

where, $a : \mathbb{R}^d \to \mathbb{R}^d$ is some function that maps the true latents to the predicted latents.

**Assumption 3.3.** Diverse perturbations: The perturbations used in the DGP as described in assumption 3.1 are diverse such that $\dim\big(\text{span}\big(\Delta\big)\big) = d$.

**Proposition 3.4.** *If assumption 3.1 and* **??** *hold, then* $dim\big(span\big(\Upsilon\big)\big) \leq d$.

As the functions described so far do not have any linearity constraints.

**Assumption 3.5.** For each component $i \in \{1, 2, ..., d\}$ of function $a$ and each component $j \in \{1, 2, ..., d\}$ of $z \in Z$, we define the set:

$$S^{ij} = \{\beta|\nabla_j a_i(z+b) = \nabla_j a_i(z) + \nabla_j^2 a_i(\beta)b, \forall z \in Z\}$$

where $b \in \mathbb{R}^d$ is a fixed vector. Each set $S^{ij}$ has a non-zero Lebesgue measure in $\mathbb{R}^d$.

If the feature extractor $f$ is constraint to be analytic, then $a$ is also analytic as $g$ is analytic.

**Theorem 3.6.** *If assumptions 3.1, 3.2, 3.3, and 3.5 hold, then the feature extractor that optimizes the loss in equation 3 identifies true latents upto an affine transformation:*

$$\hat{z} = Az + C$$

*where* $A \in \mathbb{R}^{d\times d}$ *and* $C \in \mathbb{R}^d$.

*Proof.* The proof is provided in the appendix. □

**Assumption 3.7.** We assume that each $i \in [1, 2, ..., m]$, the perturbation $\delta_i$ is small relative to $z_i$ and the excess noise generated by the feature extractor $f$ is small relative to the estimated latent variable $\hat{z}_i$. Specifically:

$$\begin{aligned} ||\delta_i|| << ||z_i|| \qquad \text{and,} \\ ||\upsilon_i|| << ||\hat{z}_i|| \end{aligned}$$

**Proposition 3.8.** *If assumptions 3.1, 3.3 and 3.7 hold true, then* $\Upsilon$ *and* $\Delta$ *are related as:*

$$\Upsilon \approx A\Delta \tag{5}$$

*Proof.* The proof is provided in the appendix. □

As we are trying to minimize $\Upsilon$, notice that $A$ being a $d \times d$ zero matrix trivially regresses $\Upsilon$ to zero. However, in that case, using theorem 3.6: $\hat{z} = A(z) + C = C$, that is, the oracle classifier $c^*$ always gets same constant input (and thus produces the same output), contradicting assumption 3.2. Therefore, we conclude that A cannot be a zero matrix with a non-trivial classifier. Additionally:

$$\begin{aligned} &\text{rank}\big(\Upsilon\big) \leq \min\Big(\text{rank}\big(A\big), \text{rank}\big(\Delta\big)\Big) \\ &\text{rank}\big(A\big) \leq d \quad \big(\text{Using 3.3 and 3.4}\big) \end{aligned} \tag{6}$$

At this point, we have a feature extractor that learns an affine transformation of the true latents and a classifier that can correctly map the true latents to the corresponding labels. Next, we define the inverse affine transformation as

$$z = A^{-1}(\hat{z} - C) \tag{7}$$

However, invertibility of $A$ cannot be guaranteed as $A$ can be a rank deficient matrix (6). Let $\hat{A}$ be a transformation function that represents the learned estimate of pseudo inverse $A^+ = (A^T A)^{-1} A^T$ using parameter $\theta_t$.

**Definition 3.9. Spurious correlations:** Let $z_i$ represent the $i^{\text{th}}$ element of $z$. If:

$$\left| \mathbb{E}_{z \in \tilde{D}} \Big[ P\Big(y|z_i\Big) \Big] - \mathbb{E}_{z \in D} \Big[ P\Big(y|z_i\Big) \Big] \right| > 0$$

then the difference between the two expectations is the spurious correlation of $y$ with $z_i$ in $\tilde{D}$.

**Definition 3.10. Spurious ratio:** Let $\mathcal{I}$ and $\mathcal{H}$ be the mutual-information and entropy functions respectively. Additionally $J(y, z_i) = \frac{\mathcal{I}(y,z_i)}{\mathcal{H}(y|z_i)}$ be the information gain ratio. We define the set of spurious ratios of the observed dataset as $\mathcal{B} = [\beta_1, \beta_2, ...\beta_d]$, where:

$$\beta_i = \frac{\left| \mathbb{E}_{z \sim \tilde{D}} \Big[ J\big(y, z_i\big) \Big] - \mathbb{E}_{z \sim D} \Big[ J\big(y, z_i\big) \Big] \right|}{\mathbb{E}_{z \sim D} \Big[ J\big(y, z_i\big) \Big]} \tag{8}$$

Additionally, we extend this definition to the learned spurious ratios by the learned model as:

$$\hat{\beta}_i = \frac{\left| \mathbb{E}_{z \sim \tilde{D}} \Big[ J\big(\hat{y}, z_i\big) \Big] - \mathbb{E}_{z \sim D} \Big[ J\big(\hat{y}, z_i\big) \Big] \right|}{\mathbb{E}_{z \sim D} \Big[ J\big(\hat{y}, z_i\big) \Big]} \tag{9}$$

**Definition 3.11. Activation variability of $A^+$:** Let $\mathcal{K}$ be a $d \times d$ matrix where each of it's elements $\kappa_{ij}$ is calculated as a relative variance as:

$$\kappa_{ij} = \frac{\sigma^2(\hat{A}_{ij} \tilde{Z}_i)}{(\hat{A}_{ij} \tilde{Z}_i)} \tag{10}$$

Here, $\sigma^2$ is the variance and *x-bar* is the mean function. We call $\kappa_{ij}$ as the activation variability of $A^+_{ij}$.

**Definition 3.12. Minimal sufficient parameterization for $A^+$:** We call a set of parameters $\theta$ as minimal sufficient parameters if $\theta$ is the smallest set of parameters that can be used to represent the pseudo-inverse $A^+$.

**Theorem 3.13.** *Under minimal sufficient parameterization for $A^+$, and if assumption 4.1 holds, then $\beta_i$ is monotonically decreasing in $\kappa_{ij}$.*

*Proof.* The proof is provided in the appendix. □

Theorem 3.13 provides a valuable insight into how activation variability inversely correlates with spurious correlations, suggesting an impactful strategy for neural network architecture and pruning design. For neural networks, pruning methods that selectively enhance activation variability — such as removing nodes or weights with low activation variance — can directly mitigate spurious correlations and improve generalization. This insight implies that pruning should specifically target "low-variance" weights or nodes, which are more likely to contribute to high spurious correlations in learned representations. This insight align as well as extend the empirical evidence and analyses from plethora of prior works, which shows that sparse networks retaining only the most effective parameters help reduce spurious correlations.

## 4. Pruning for robust discovery of latents

In practical settings, we lack direct access to the true data-generating process and instead rely on sampled datasets that may fail to accurately reflect the underlying mechanisms. We identify two primary sources of mis-specification within the observed data distribution: (1) spurious correlations, where the data contains relationships that do not hold causally or robustly across environments, and (2) mislabeled data, where errors in the labeling process introduce noise that diverges from the true data-generation process. These two mechanisms represent distinct forms of weak supervision that can impair model performance and generalization. Formally, we define the following assumptions to characterize them:

**Definition 4.1. Spurious observed distribution**: Let $\tilde{D}_{\text{spur}} \subset D$ be an observed data-distribution such that it has $m$ number of spurious correlations of $y$ with $z_i \in \tilde{D}_{\text{spur}}, \forall i \in \{1, 2, ..., d\}$.

**Definition 4.2. Mislabelled DGP**: Let the DGP described in assumption 3.1 be modified as follows:

$$y \leftarrow h(z + \eta), \quad \eta \sim \mathcal{N}$$

where $\eta$ is the additive random noise sampled from some distribution. Additionally, we assume

$$\forall z \in \tilde{Z} : \sum_i \phi\Big(h(z|_{z_i = z_i + \eta_i}) \neq h(z)\Big) = m.$$

Here, $i \in [1, 2, 3, ..., d]$ represent the $i^{\text{th}}$ element of $z \in \tilde{Z}$ and $\phi$ is the Kronecker Delta function.

**Assumption 4.3. Sufficient coverage**: For both definitions 4.1 and 4.2, we ensures sufficient coverage of the true data-generating process despite spurious correlations or label noise by assuming:

$$m \ll d \times |Z| \tag{11}$$

**Proposition 4.4.** *Let $\tilde{D}_{miss}$ denote the observed variable-label pairs generated from the DGP described in Assumption 4.2. If Assumption 4.3 holds, then:*

$$\mathbb{E}_{(z,y) \sim \tilde{D}_{spur}} \Big[ \mathcal{I}\big(y|z\big) \Big] = \mathbb{E}_{(z,y) \sim \tilde{D}_{miss}} \Big[ \mathcal{I}\big(y|z\big) \Big] \tag{12}$$

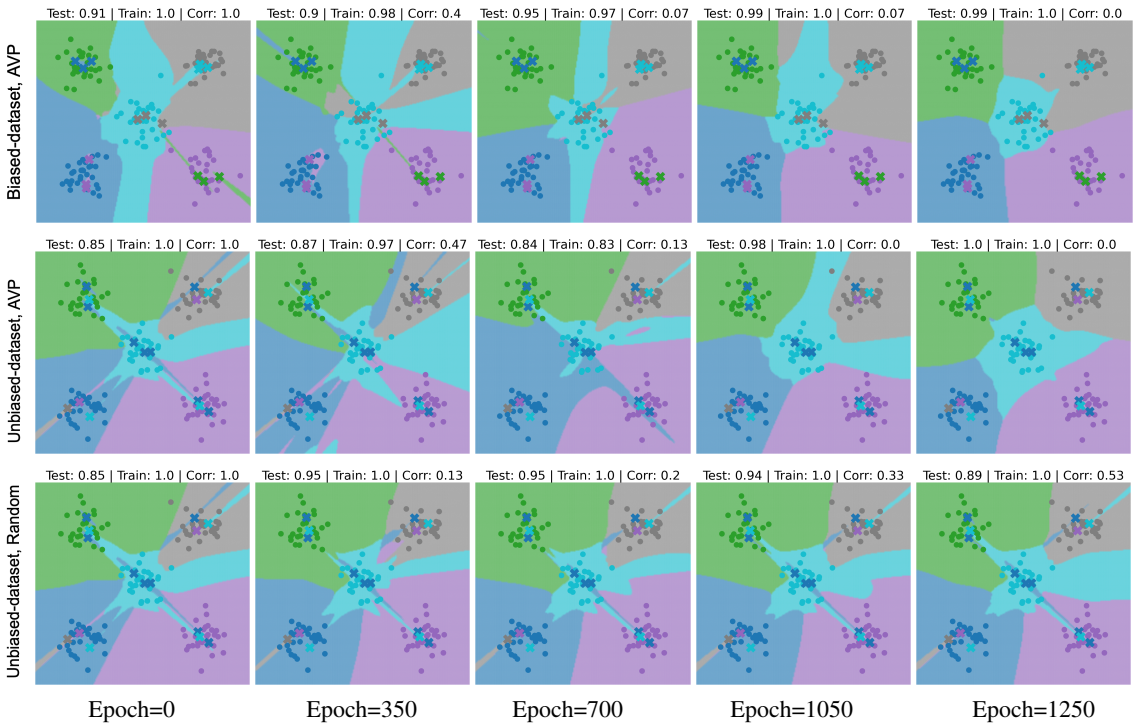


*Figure 1.* The figure shows decision boundary evolution across epochs for AVP on biased and unbiased datasets, and random pruning on the unbiased dataset. Testing points (circles) are colored by true class, while corrupted points (crosses) are colored by corrupted labels.

*Proof.* The proof is provided in the appendix. □

**Corollary 4.5.** *Let $\tilde{D}_{spur}$ the observed variable-label pairs and described in 4.1 and $\tilde{D}_{miss}$ be the observed variable-label pairs from DGP in 4.2, then the spurious ratios of both datasets are equal. i.e, $\mathcal{B}_{\tilde{D}_{spur}} = \mathcal{B}_{\tilde{D}_{miss}}$*

*Proof.* The result follows immediately from 4.4. □

Corollary 4.5 reveals that the spurious ratio $\beta_i$ remains invariant, whether weak supervision arises from the bias in the data-generation process ($\tilde{D}_{\text{spur}}$) or variances introduced during sampling and labeling ($\tilde{D}_{\text{miss}}$). This invariance implies that pruning strategies based on activation variability can consistently reduce spurious correlations, regardless of their origin. This universality simplifies the approach to addressing weak supervision, providing a unified framework for mitigating spurious correlations across diverse scenarios.

### 4.1. Theory-Informed Pruning

The theoretical results presented above rely on strong assumptions, such as minimal sufficient parameterization, which can be difficult to achieve in practical scenarios. These assumptions, while idealized, provide valuable insights and guidance for developing effective methodologies. Based on these insights, we propose and then test the following hypothesis:

***Hypothesis**: Pruning strategies guided by activation variability can effectively isolate robust features, thereby enhancing model generalization.*

We design experiments to evaluate whether pruning based on activation variability replicates the theoretical benefits under practical conditions, focusing on its impact on model generalization and feature robustness. These experiments aim to validate the hypothesis that such pruning effectively isolates robust features and mitigates the effects of weak supervision caused by spurious correlations or labeling noise. By simulating these biases in the data-generation process and sampling or labeling mechanisms, we assess the ability of the proposed pruning strategy to address dataset imperfections and enhance model performance.

**Experimental Setup:** In our experiments, we construct a synthetic classification problem with five classes. Each class is modeled as a 2D Gaussian distribution, where the means are uniformly distributed around a unit circle, with an additional class centered at the origin. All classes share the same variance. We generate two types of datasets to simulate weak supervision:

1. *Biased Dataset*: In this dataset, samples from class $i$ are

consistently mislabeled as a fixed class $j$, where the class $j$ is unique for each class $i$. This introduces a structured form of mislabeling across the dataset.

2. *Unbiased Dataset*: In this dataset, samples from class $i$ are randomly mislabeled as any other class. Unlike the biased dataset, the mislabeling is unstructured and varies at the sample level.

**Training details**: We train a fully connected neural network with two hidden layers, each with 32 neurons and *Tanh* activation functions. The input is the 2D coordinate $(x, y)$ of each point, and the output represents unweighted class probabilities. The network is trained using the cross-entropy loss function and the Adam optimizer. Due to the over-parameterized nature of the network, we train it on each dataset until it achieves zero training error, capturing both robust and spurious features.

Over-fitting the model before pruning allows us to evaluate how pruning strategies affect the network's ability to retain robust features while eliminating reliance on spurious patterns. Two pruning strategies are compared: Activation variability-based pruning (our method) and Random-pruning. Activation variability-based pruning selects parameters with the lowest activation variability for removal, prioritizing robust features. Random-pruning, by contrast, removes parameters at random, serving as a baseline. For both methods, we prune two candidates per epoch with a probability of 0.5 and evaluate performance and evolution of the decision boundary over 1200 epochs.

**Results**: We can draw the following conclusions from the results of our experiments shown in Fig 1:

- AVP consistently produces decision boundaries that align closely with the true class labels for both biased and unbiased datasets. This ability to refine boundaries progressively, despite the presence of corrupted samples, highlights its effectiveness in isolating robust features and reducing reliance on spurious correlations.
- AVP's performance on both the biased and unbiased dataset confirms our hypothesis that activation variability-based pruning can effectively isolate robust features and enhancing model generalization under weakly supervised conditions irrespective of the weather weak supervision arises from structural biases in the data-generation process or noise in data sampling and labeling.
- Random pruning fails to achieve similar results, producing irregular and unstable decision boundaries. However, it should be noted that even with random pruning, the model's generalization performance improves compared to the original over-fitted model. This is in line with the prior literature indicating that pruning itself is beneficial for robustness of predictions.

# 5. Experiments

## 5.1. Experimental Setup

Our preliminary experiments demonstrated the effectiveness of our method in identifying and pruning latent variables to improve robustness and generalization in a simplistic setting. In this section, we scale up our evaluation to real-world scenarios, comparing our approach against state-of-the-art methods and other baselines. We benchmark our approach against state-of-the-art methods and other baselines, including Empirical Risk Minimization (ERM), Random features pruning, Group Distributionally Robust Optimization (Group DRO), and Group-Conditional DRO (GC-DRO), as described in (Zhou et al., 2021). Experiment details are provided below.

### 5.1.1. Datasets, Tasks and Implementations

We follow the experimental setup outlined in (Zhou et al., 2021) to ensure a fair comparison. We evaluate our method with prior works on three datasets: CelebA, MNLI, and FDCL18, each representing a different domain and task. We create clean and imperfect partitions for each dataset to evaluate the robustness of the models against spurious correlations. Details about the datasets and tasks are provided below, with more information available in the appendix.

- **CelebA (Object Recognition):** The dataset consists of 162,770 images of celebrity faces and is used for classifying hair color (blonde or dark). Gender serves as a spurious attribute, with "male" disproportionately associated with "dark hair" and "female" with "blonde hair." For evaluation, we create clean partitions with four groups based on combinations of hair color and gender, and imperfect partitions where certain combinations (e.g., "dark-haired males") are overrepresented. ResNet-18 pretrained on ImageNet is fine-tuned with Adam for 50 epochs.
- 
- **MNLI (Natural Language Inference):** The dataset contains 206,175 sentence pairs tasked with determining whether a hypothesis is true (entailment), false (contradiction) or undetermined (neutral) given a premise. Negation words (e.g., "nobody,"nothing") serve as spurious features, often associated with the "contradiction" label. Clean partitions divide the dataset into nine groups based on negation levels and labels, while imperfect partitions mix negation features across three groups. *RoBERTa* is fine-tuned using Adam optimizer for this task.
- **FDCL18 (Toxicity Detection):** This dataset includes 100,000 tweets labeled as hateful, spam, abusive, or normal, with dialect variations acting as spurious attributes correlated with toxicity labels. Clean partitions divide

*Table 1.* Table caption

| Dataset | ERM | RS | G-DRO (EG) | G-DRO(Greedy) | GC-DRO | R-prune | AVP (our) |
|---|---|---|---|---|---|---|---|
| CelebA (clean) | 40.14 | 86.81 | 88.19 | 88.19 | 88.75 | - | - |
| CelebA (imperfect) | 40.14 | 44.17 | 45.97 | 45.97 | 82.85 | - | - |
| MNLI (clean) | 70.84 | 67.02 | 75.14 | 75.14 | 77.82 | - | - |
| MNLI (imperfect) | 70.84 | 67.26 | 70.34 | 70.34 | 75.32 | - | - |
| FDCL18 (clean) | 34.30 | 55.44 | 56.83 | 56.83 | 57.28 | - | - |
| FDCL18 (imperfect) | 34.30 | 26.10 | 36.24 | 36.24 | 48.42 | - | - |

the data into 16 groups based on combinations of dialects and toxicity labels, while imperfect partitions are grouped solely by dialect. Again, *RoBERTa* is fine-tuned using Adam for this task.

We use two evaluation metrics to compare the performance of our method with other baselines: robust accuracy and average accuracy. Robust accuracy is the minimum accuracy across all groups in the clean test partitions, while average accuracy is the overall accuracy across the test data. Both metrics are averaged over five independent runs with random seeds.

**Methods Compared** We compare the following approaches for our experiments:

- **Empirical Risk Minimization (ERM):** Standard approach that minimizes the average training loss, ignoring spurious correlations or group-level biases.
- **Resampling (RS):** This method rebalances data by oversampling underrepresented groups during training to achieve uniform group representation.
- **Group distributionally robust optimization (G-DRO):** Refer to the original work (Zhou et al., 2021) for more details about G-DRO-EG G-DRO-Greedy, and GC-DRO.
- **Random Features Pruning (R-prune):** Randomly removes a fixed percentage of features each epoch to mitigate spurious correlations.
- **Our Method:** Frequency-based pruning identifies and removes low-variance activations to mitigate spurious correlations while preserving robust features.

### 5.2. Results

The robust and average accuracies for all methods across datasets and partitions are summarized in Table 1.

## References


Ahuja, K., Hartford, J. S., and Bengio, Y. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35: 15516–15528, 2022.

Chen, T., Zhang, Z., Wang, P., Balachandra, S., Ma, H., Wang, Z., and Wang, Z. Sparsity winning twice: Better robust generalization from more efficient training. *arXiv preprint arXiv:2202.09844*, 2022.

He, Z., Xie, Z., Zhu, Q., and Qin, Z. Sparse double descent: Where network pruning aggravates overfitting. In *International Conference on Machine Learning*, pp. 8635–8659. PMLR, 2022.

Hoefler, T., Alistarh, D., Ben-Nun, T., Dryden, N., and Peste, A. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

Mityagin, B. The zero set of a real analytic function. *arXiv preprint arXiv:1512.07276*, 2015.

Nagarajan, V. and Kolter, J. Z. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Tripuraneni, N., Adlam, B., and Pennington, J. Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, 34:13883–13897, 2021.

Victor, V., D'Amour Alexander, Y. S., and Jacob, E. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv: 2106.00545*, 2021.

Zhou, C., Ma, X., Michel, P., and Neubig, G. Examining and combating spurious features under distribution shift. In *International Conference on Machine Learning*, pp. 12857–12867. PMLR, 2021.

# A. Appendix section name

**Theorem A.1.** *If assumptions 3.1, 3.2,* **??***, and 3.5 hold, then the feature extractor that solves equation 3 identifies true latents upto an affine transformation:*

$$\hat{z} = Az + C$$

*where* $A \in \mathbb{R}^{d \times d}$ *and* $C \in \mathbb{R}^d$.

*Proof.* From equation 2 we have:

$$f(x_i) + \delta_i^{'} = f(\tilde{x}_i) \tag{13}$$
$$f \circ g(z_i) + \delta_i^{'} = f \circ g(\tilde{(z)}_i) \tag{14}$$
$$a(z_i) + \delta_i^{'} = a(\tilde{z}_i) \tag{15}$$
$$a(z_i) + \delta_i^{'} = a(z_i + \delta_i) \tag{16}$$

Denoting the $j^{\text{th}}$ component of $a(z_i + \delta_i)$ as $a_j(z_i + \delta_i)$, and taking the gradients of the RHS:

$$\nabla_{z_i} a_j(z_i + \delta_i) = \left(\frac{du_i}{dz}\right)^T \nabla_{u_i} a_j(u_i), \quad \text{where, } u_i = z_i + \delta_i$$

$\left(\frac{du_i}{dz}\right)$ is the Jacobian of $u_i$ with respect to $z_i$, and it equates to identity. So the above equation becomes:

$$\nabla_{z_i} a_j(z_i + \delta_i) = \nabla_{u_i} a_j(u_i)$$

Repeating the above for all $j \in \{1, 2, ..., d\}$, we get:

$$\begin{aligned}\left[\nabla_{z_i} a_1(z_i + \delta_i) + ... + \nabla_{z_i} a_d(z_i + \delta_i)\right] &= \left[\nabla_{u_i} a_1(z_i + \delta_i) + ... + \nabla_{u_i} a_d(z_i + \delta_i)\right] \\ &= J^T(z_i + \delta_i)\end{aligned}$$

Here, $J^T$ is the Jacobian of $a$ computed at $z_i + \delta_i$. Now, using 16, we get:

$$a(z_i + \delta_i) = a(z_i) + \delta_i^{'} \tag{17}$$
$$J^T(z + \delta_i) - J^T(z) = 0 \tag{18}$$

Considering the $j^{\text{th}}$ component of the above equation, we get:

$$\begin{aligned}&\nabla a_j(z_i + \delta_i) - \nabla a_j(z_i) = 0 \\ &\equiv \begin{bmatrix} \nabla_1^2 a_j(\beta_{1j}) \\ \nabla_2^2 a_j(\beta_{2j}) \\ \cdot \\ \cdot \\ \cdot \\ \nabla_d^2 a_j(\beta_{dj}) \end{bmatrix} \delta_i = 0\end{aligned}$$

Here, $\nabla^2$ is the Hessian and $\nabla_i^2 a_j(\beta_{ij})$ is the $i^{\text{th}}$ column and the $j^{\text{th}}$ row of the Hessian matrix. From assumption 3.5, it follows that $\nabla_i^2 a_j(\beta_{ij})\delta_i = 0$ over a set with non-zero Lebesgue measure. Since, $a_j$ is analytic, we can conclude that $\nabla_i^2 a_j(z)\delta_i = 0$ for all $z$ (Mityagin, 2015). Extending the same argument over all $i$, we get:

$$\nabla^2 a_j(z)\delta_i = 0, \quad \forall j \in \{1, 2, ..., d\} \tag{19}$$
$$\nabla^2 a_j(z) = 0, \quad \forall j \in \{1, 2, ..., d\} \quad \text{as } \Delta = \{\delta_i\}_{\forall i} \quad \text{is a linearly independent set} \tag{20}$$

Equation 20 implies that the Hessian of $a$ is a zero matrix. Since $a$ is analytic, it follows that $a$ is an affine function. Hence, we can write $a(z) = Az + C$ for some $A \in \mathbb{R}^{d\times d}$ and $C \in \mathbb{R}^d$, completing the proof. □

**Theorem A.2.** *If assumptions 3.1, 3.3 and 3.7 hold true, then $\Upsilon$ and $\Delta$ are related as:*

$$\Upsilon \approx Q\Delta \tag{21}$$

*Here, $Q$ is a jacobian matrix that represents the local linear transformation around $z$.*

*Proof.* Given the perturbed latent variable $z + \delta_i$, the feature extractor's output $f(\tilde{x}_i) = f\big(g(z + \delta_i)\big)$ can be approximated using a first order taylor expansion around $z$, as $\delta_i$ is small (assumption 3.7):

$$f(\tilde{x}_i) = f\big(g(z + \delta_i)\big) \approx f(g(z)) + \nabla f\big(g(z)\big)\delta_i$$

where $\nabla f\big(g(z)\big)$ is the Jacobian of $f$ with respect to $z$ at $g(z)$. Let $Q = \nabla f\big(g(z)\big)$, which is a constant matrix under the small perturbation assumption. Then the feature extractor's output becomes:

$$f(\tilde{x}_i) = f\big(g(z + \delta_i)\big) \approx f(g(z)) + Q\delta_i$$

Defining the feature extractor's noise $\upsilon_i$ as the difference between the perturbed and unperturbed outputs of the feature extractor:

$$\upsilon_i = f(\tilde{x}_i) - f(g(z)) \approx Q\delta_i$$

Since, this holds for all $\delta_i \in \Delta$, we can aggregate as:

$$\begin{aligned} &\Upsilon \approx \mathbb{E}_{z\sim\mathbb{P}_z}[Q]\Delta. \\ \equiv\ &\Upsilon \approx A\Delta. \end{aligned}$$

Thus $\Upsilon$ can be approximately expressed as a linear transformation of the perturbations $\Delta$ through the matrix $A$, completing the proof. □

**Theorem A.3.** *Under minimal sufficient parameterization for $A^+$, and if If assumption 4.1 holds, $\alpha_i$ is monotonically decreasing in $\mathcal{A}_{ij}$.*

*Proof.* The proof can be done using the following statements.

- Higher $\mathcal{A}_{ij}$ values indicate a higher variance to mean ratio in $\frac{\sigma^2(\hat{A}_{ij}\tilde{Z}_i)}{(\hat{A}_{ij}\tilde{Z}_i)}$, suggesting a more pronounced and consistent effect of $\tilde{A}_{ij}$ on $\tilde{Z}_j$ (assumption 7) and consequently $Z_j$ (assumption 4.1).
- With an increase in $\mathcal{A}_{ij}$, the difference between $\mathbb{E}_{z\sim\tilde{D}}[J(y, z_i)]$ and $\mathbb{E}_{z\sim D}[J(y, z_i)]$ decreases. This is because the effect of added noise in $\tilde{Z}$ becomes less significant over the more pronounced effect of $\tilde{A}_{ij}$.
- Since the denominator $\mathbb{E}_{z\sim D}[J(y, z_i)]$ remains constant in 8, larger $\mathcal{A}_{ij}$ imply smaller $\beta_i$. Thus finishing the proof.

□

**Proposition A.4.** *Let $\tilde{D}_{miss}$ denote the observed variable-label pairs generated from the DGP described in Assumption 4.2. If Assumption 4.3 holds, then:*

$$\mathbb{E}_{(z,y)\sim\tilde{D}_{spur}}\Big[\mathcal{I}\big(y|z\big)\Big] = \mathbb{E}_{(z,y)\sim\tilde{D}_{miss}}\Big[\mathcal{I}\big(y|z\big)\Big] \tag{22}$$

*Proof.* We have,

$$\begin{aligned} \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{I}\big(y|z\big)\Big] &= \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{H}(y) - \mathcal{H}\big(y|z\big)\Big] \\ &= \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{H}(y)\Big] - \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{H}\big(y|z\big)\Big] \\ \text{similarly,}\quad \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{I}\big(y|z\big)\Big] &= \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{H}(y)\Big] - \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{H}\big(y|z\big)\Big] \end{aligned}$$

Since $\mathcal{H}(y)$ depends only on the marginal distribution of $y$, and both $\tilde{D}_{\text{spur}}$ and $\tilde{D}_{\text{miss}}$ share the same marginal distribution of $y$, we have:

$$\mathbb{E}_{(y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{H}(y)\Big] = \mathbb{E}_{(y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{H}(y)\Big]$$

Using definitions 4.1 (spurious correlations) and 4.2 (label noise), we have:

$$\mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{H}(y|z)\Big] = \mathbb{E}_{(z,y)\sim D}\Big[\mathcal{H}(y|z)\Big] + \mathcal{O}\bigg(\frac{m}{d\times|Z|}\bigg)$$

$$\text{and,} \qquad \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{H}(y|z)\Big] = \mathbb{E}_{(z,y)\sim D}\Big[\mathcal{H}(y|z)\Big] + \mathcal{O}\bigg(\frac{m}{d\times|Z|}\bigg)$$

Using assumption 4.3 and the above equations, we get $\mathbb{E}_{(z,y)\sim\tilde{D}_{\text{spur}}}\Big[\mathcal{I}\big(y|z\big)\Big] = \mathbb{E}_{(z,y)\sim\tilde{D}_{\text{miss}}}\Big[\mathcal{I}\big(y|z\big)\Big]$, thus concluding the proof.

□

### A.1. Datasets, tasks and implementation details

**CelebA (Object Recognition)** The CelebA dataset consists of 162,770 celebrity face images, where the task is to predict hair color (blonde or dark). A significant challenge arises from the spurious correlation between gender (male or female) and hair color, as males are predominantly associated with dark hair and females with blonde hair. We define two partitioning schemes: a clean partition with four groups based on all combinations of hair color and gender, and an imperfect partition with only two groups, where spurious correlations dominate. For instance, "dark-haired males" are overrepresented, while "blonde-haired males" are rare. A ResNet-18 pretrained on ImageNet is fine-tuned using cross-entropy loss. We use the Adam optimizer with a learning rate of x, weight decay of y, and batch size of 128 for 50 epochs. Frequency-based pruning is applied from epoch 10, progressively targeting low-variance activations.

**MNLI (Natural Language Inference)** The MultiNLI dataset contains 206,175 premise-hypothesis pairs, with the task of classifying their relationship as entailment, contradiction, or neutral. Spurious correlations emerge due to the frequent association of negation words (e.g., "nobody," "nothing") with the contradiction label. We create a clean partition with nine groups based on three negation levels (no negation, negation group 1, negation group 2) and three labels, while the imperfect partition reduces this to three groups, mixing negation levels across labels to amplify spurious correlations. For implementation, we fine-tune a pretrained RoBERTa model with a learning rate of x and batch size 32. Hyperparameters for imperfect partitions include $\alpha = 0.5$ and $\beta = 0.2$. Frequency-based pruning is applied to intermediate layers, prioritizing activations with high variability to retain robust features.

**FDCL18 (Toxicity Detection)** The FDCL18 dataset contains 100,000 tweets annotated with four toxicity labels: hateful, spam, abusive, or normal. The spurious attribute in this dataset is dialect variation, with African American English (AAE), White-aligned English, and other dialects spuriously correlated with toxicity labels, reflecting real-world biases. We use a clean partition with 16 groups combining dialects and toxicity labels and an imperfect partition with four groups based on dialect alone, mixing toxicity labels within each group. A pretrained RoBERTa model is fine-tuned with a learning rate of x and batch size 32. For imperfect partitions, hyperparameters are set to $\alpha = 0.5$ and $\beta = 0.25$. Frequency-based pruning focuses on latent feature representations, ensuring stability across dialect groups while suppressing spurious correlations.