Theory of Agreement-on-the-Line in Linear Models and Gaussian Data

Christina Baek
Carnegie Mellon University
kbaek@cs.cmu.edu

Aditi Raghunthan
Carnegie Mellon University
aditirag@cs.cmu.edu

Zico Kolter
Carnegie Mellon University
zkolter@cs.cmu.edu

Abstract

Under distribution shifts, deep networks exhibit a surprising phenomenon: in-distribution (ID) versus out-of-distribution (OOD) accuracy is often strongly linearly correlated across architectures and hyperparameters, accompanied by the same linear trend in ID versus OOD agreement between the predictions of any pair of such independently trained networks. The latter phenomenon called “agreement-on-the-line” enables precise unlabeled OOD performance estimation of models. In this work, we discover that agreement-on-the-line emerges even in linear classifiers over Gaussian class conditional distributions. We provide theoretical guarantees for this phenomenon in classifiers optimized via randomly initialized gradient descent, approximated by linear interpolations between random vectors and the Bayes-optimal classifier. Next, we prove a lower bound on the residual of the correlation between ID versus OOD agreement that grows proportionally with the residual of accuracy. Real-world experiments on CIFAR10C shifts, validate our findings and the broader relevance of our theoretical framework.

1 INTRODUCTION

A long standing challenge with deep neural networks is their tendency to perform unreliably under distribution shifts. Surprisingly, a substantial collection of recent works show that the model performance of neural networks on natural image and language shifts tends to degrade in a highly predictable fashion (Miller et al., 2021, 2020; Shankar et al., 2020). In particular, many real-world benchmarks observe a phenomenon called “accuracy-on-the-line” (Miller et al., 2021) where the in-distribution (ID) and out-of-distribution (OOD) classification accuracies of deep networks are strongly linearly correlated across architectures and hyperparameters, as measured by the coefficient of determination (R^2).

Following this observation, Baek et al. (2022) discovered a coupled phenomenon called “agreement-on-the-line” — in circumstances where a set of deep networks observe strong correlation in ID versus OOD accuracy, the *agreement* (the rate at which the predictions of two models agree) between these models is also strongly linearly correlated ID versus OOD with the *same slope and bias*. On the other hand, when the linear correlation of accuracy is weak, the linear correlation of agreement is also weak[1]. Accuracy-on-the-line and agreement-on-the-line together provide a neat paradigm for OOD performance estimation without any labeled data. When accuracy-on-the-line holds, OOD performance is a univariate function of ID performance alone, independent of any algorithmic choices made during neural network training. Agreement-on-the-line is useful in data constrained regimes because agreement is a not a label-dependent quantity. First, without any OOD labels, one can verify if accuracy-on-the-line holds by checking for agreement-on-the-line. Second, since the linear fit of these trends match, one can simply transform ID accuracies by the slope and bias of agreement-on-the-line to get a close estimate of OOD accuracies.

While these phenomena have immediate practical value, the current literature lacks theory that explains *why agreement-on-the-line occurs jointly with*

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

[1] “Agreement-on-the-line” often refers to the strongly and weakly correlated cases jointly. However, we will use this term to only refer to the prior case when ID versus OOD agreement is strongly correl ated with the same slope and bias as accuracy-on-the-line.

accuracy-on-the-line. Understanding this question is critical to guarantee when agreement-on-the-line provides accurate estimates of OOD performances in practice. Although several works provide theoretical guarantees for the accuracy-on-the-line phenomenon (LeJeune et al., 2024; Mania and Sra, 2020), existing analyses of agreement-on-the-line has been limited to settings where the ID versus OOD trends of accuracy and agreement do not match in both slope and intercept (Lee et al., 2023). Further complicating theoretical analysis, previous empirical findings had suggested that agreement-on-the-line is a phenomenon specific to deep networks and not simpler model families, e.g. linear classifiers (Baek et al., 2022).

In this work, we establish formal guarantees for agreement-on-the-line in a simple setting of high-dimensional linear classifiers and Gaussian class conditional distributions. In particular, we analyze sets of linear models formed by taking convex combinations of random vectors and the optimal Bayes classifier – a construct we call the *convex collection*. Surprisingly, both convex collections and sets of linear models optimized by randomly initialized gradient descent exhibit agreement-on-the-line and accuracy-on-the-line under our data setup. Moreover, our theoretical findings closely predict when agreement-on-the-line emerges in deep neural networks on real benchmarks. A detailed summary of our contributions is outlined below.

- In §4, we characterize conditions under which agreement-on-the-line occurs in convex collections. We formalize the measure called the *similarity score* between pairs of models, and characterize the "good range" of similarity scores where ID versus OOD agreement is strongly correlated with matching linear trend as accuracy-on-the-line. We prove when model pairs in the convex collection fall in this good range, depending on the degree of distribution shift, learnability of the ID task, and the eigenvalue decay rate of the class-conditioned covariance matrices.
- In §5, we prove a lower bound on the absolute residual of agreement-on-the-line trend that grows proportionally to that of the accuracy-on-the-line trend. This guarantees the absence of the false positive event where agreement-on-the-line occurs but accuracy-on-the-line does not.
- In §6, we validate our findings from §4 and §5 on CIFAR10C (Hendrycks and Dietterich, 2019) in deep networks and CLIP (Radford et al., 2021). Our conclusions in linear models well characterize when agreement-on-the-line occurs in deep networks if we measure how the shift is encoded in the penultimate representation space.

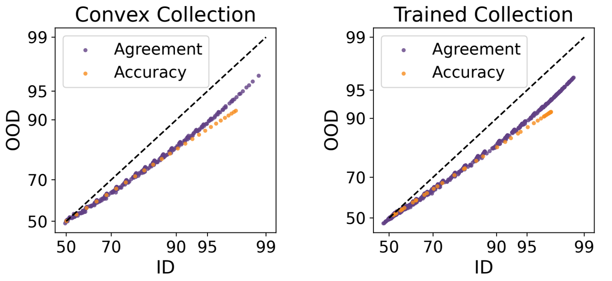


Figure 1: **Convex and trained collections** Both convex (Left) and SGD trained (Right) model collections observe accuracy-on-the-line and agreement-on-the-line under scale shifts, where the class-conditioned covariance matrix shifts from $\Sigma = {}^{1}/_{2} \cdot I \to I$.

2 RELATED WORKS

Accuracy-on-the-Line Accuracy-on-the-line is a widely occurring phenomenon in deep networks across distribution shift benchmarks such as regional and temporal shifts in the WILDS dataset (Koh et al., 2021), corrupted ImageNet and CIFAR10 (Hendrycks and Dietterich, 2019), dataset reproductions of ImageNet and MNIST (Recht et al., 2019; Shankar et al., 2020; Yadav and Bottou, 2019), NLP tasks such as text classification and question-answering (Miller et al., 2020; Awadalla et al., 2022), and Kaggle competition train-test splits (Roelofs et al., 2019). This strong ID versus OOD correlation in accuracy has been shown to hold across models of various architectures, hyperparameters, and training duration (Miller et al., 2021).

Several works have tried to uncover the theoretical underpinnings of accuracy-on-the-line. Mania and Sra (2020) observe the phenomenona under assumptions on how models learn: the set of in-distribution examples where weaker models predict correctly must not overlap with examples that stronger models predict incorrectly. Similar to our work, others have studied the problem in linear models and Gaussian data. In particular, we study distribution shifts over symmetric Gaussian class-conditional distributions and binary class labels Miller et al. (2021). They showed that only under distribution shifts that simply scale the norm of the class mean or covariance by a constant factor, any arbitrary collection of linear classifier would observe perfect accuracy-on-the-line with $R^2 = 1$. For shifts that further change the *direction or shape* of the class mean or the covariance matrix, accuracy-on-the-line occurs under specific asymptotic conditions (Miller et al., 2021; LeJeune et al., 2024). In our work, we show that agreement-on-the-line requires additional conditions beyond those required for accuracy-on-the-line.

Agreement-on-the-line Baek et al. (2022) observed that under tasks where accuracy-on-the-line holds, a similar phenomenon also holds for the agreement between pairs of neural network classifiers: the ID versus OOD agreement between the predictions of any two pairs of neural network classifiers also observes a strong linear correlation with the same slope and bias as the linear fit of accuracy-on-the-line. Recently, Saxena et al. (2024) extend these findings to lightly-finetuned foundation models. Of particular relevance to our work, they observe that even collections of linear classifiers finetuned from random initialization on top of frozen embeddings can observe on-the-line trends.

Previously, Lee et al. (2023) analyzed agreement-on-the-line in 2-layer deep linear networks evaluated by mean squared error. Similar to our work, they prove that models trained from randomly initialized weights exhibit agreement-on-the-line. However, they were only able to observe matching slopes and not matching intercepts between the linear fit of accuracy and agreement. In our work, we return to the classification setting in linear models and evaluate accuracy by $0-1$ loss, consistent with previous empirical findings (Miller et al., 2021; Baek et al., 2022). We prove conditions under which the ID versus OOD accuracy and agreement linear trends match perfectly.

3 PRELIMINARIES

3.1 Notations

We use bolded lower case letters $\boldsymbol{x}$ for vectors, unbolded lower case letters x for scalars, and capital letters A for matrices. $\boldsymbol{w}$ will be used to refer to the weights of a generic linear classifier. Bolded $\boldsymbol{w}_i$ refers to the ith classifier in some model collection $\{\boldsymbol{w}_i\}_{i=0}^n$. Unbolded w_{ij} stands for the jth entry in model $\boldsymbol{w}_i$.

3.2 Data

Consider a simple binary classification problem where the examples of each class are normally distributed. First, the class label y is uniform over $\{-1, 1\}$ on both the original distribution D and the shifted distribution D'. Next, conditioned on y, the data $\boldsymbol{x} \in \mathbb{R}^d$ is normally distributed i.e.

$$\boldsymbol{x} \mid y \sim \mathcal{N}(y\boldsymbol{\mu}; \sigma^2 \Sigma_{d\times d}).$$

Without loss of generality, we fix $\boldsymbol{\mu} = \sqrt{d^{-1}}\mathbf{1}$ and $\sigma = 1/2$ in all our experiments. The shifted distribution D' can change as follows:

$$\begin{aligned}&\text{Mean Shift:} \quad \boldsymbol{\mu}' = \alpha\boldsymbol{\mu} + \beta\boldsymbol{\Delta}_\mu \\ &\text{Covariance Scale Shift:} \quad \sigma' = \gamma\sigma \ \text{ subject to } \ \gamma > 0 \\ &\text{Covariance Shape Shift:} \quad \Sigma' = \Sigma + \Delta_\Sigma\end{aligned}$$

where $\alpha, \beta > 0$ are fixed scalars and $\boldsymbol{\Delta}_\mu \in \mathbb{R}^d$ where $\|\boldsymbol{\Delta}_\mu\| = 1$ and $\langle \boldsymbol{\mu}, \boldsymbol{\Delta}_\mu \rangle = 0$.

Decay Rate of Covariance Matrix's Eigenvalues As we will observe, whether we observe accuracy-on-the-line and agreement-on-the-line depends on the shape of the covariance matrix. To study this formally, we will study the degree of accuracy-on-the-line and agreement-on-the-line under different covariance matrix shapes of the form

$$\Sigma_d^{-\alpha} = N \cdot \operatorname{diag}\left(\left[k^{-\alpha}\right]_{k=1}^d\right) \text{ where } N = \frac{d}{\sum_{i=1}^d k^{-\alpha}} \tag{1}$$

where N is a normalization constant and $\alpha \geq 0$ controls the decay rate of the eigenvalues of the covariance matrix. We refer to any covariance matrix with $\alpha \gg 0$ as "fast-decaying".

3.3 Model Collection

We consider linear classifiers $f(x) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ without bias. Accuracy-on-the-line and agreement-on-the-line is a phenomena that occurs for an *collection of trained classifiers*. In our study, we will simulate "trained" classifiers as linear combinations of the Bayes optimal classifier $\boldsymbol{w}^\star = \Sigma^{-1}\boldsymbol{\mu}/\left\|\Sigma^{-1}\boldsymbol{\mu}\right\|$ and randomly initialized weights $\boldsymbol{\epsilon}$ uniformly sampled from the set $\mathcal{W} = \{\boldsymbol{x} \mid \boldsymbol{x} \in \mathcal{S}^{d-1}, \langle \boldsymbol{w}^\star, \boldsymbol{x} \rangle = 0\}$. Specifically, consider the hypothesis distribution $\boldsymbol{w}_i \sim \mathcal{H}_{a,b}$ where

$$\boldsymbol{w}_i = a\frac{\Sigma^{-1}\boldsymbol{\mu}}{\|\Sigma^{-1}\boldsymbol{\mu}\|} + b\boldsymbol{\epsilon} \tag{2}$$

and $a^2 + b^2 = 1$ for $a, b \in [0, 1]$. Roughly, $\mathcal{H}_{a,b}$ simulates the distribution of linear models trained from random initialization for finite gradients steps. Models with a range of ID accuracies can be collected by sampling models from $\mathcal{H}_{a,b}$ induced by every value a, reaching the Bayes optimal classifier $\boldsymbol{w}^\star$ when $a = 1$. By slightly abusing the term "convex", we call such collections of models the *convex collection.* In the next paragraph, we discuss in detail the relationship between convex models and trained models.

Connection to Trained Models We note several similarities and differences between convex models and logistic models optimized by randomly initialized gradient descent. First, in our data setting with symmetric Gaussian class-conditioned distributions, the expected risk minimizer of logistic regression is equal to the optimal Bayes classifier (Bishop and Nasrabadi, 2006). Thus, in the data limit, the start and end points of the optimization trajectory is the same between trained models and convex models. However, while convex models follow a strict linear trajectory from random initialization to the optimal Bayes classifier, the training trajectory of any logistic classifier is noisy and gradient flow is not generally linear. Second, our model assumptions loosely resemble those of

LeJeune et al. (2024). Notably, they similarly study accuracy-on-the-line in linear combinations of random vectors and the optimal classifier: $w = \Sigma(a\boldsymbol{w}^{\star} + b\boldsymbol{\epsilon})$ where $\boldsymbol{\epsilon} \sim N(0, I)$. They provably show that in the proportional asymptotic regime, the regularized empirical risk minimizer follows this form under certain data assumptions different from our binary mixture model. Finally, we provide strong empirical evidence that the convex collection achieves the same ID versus OOD accuracy and agreement trends as a set of trained models, as shown in Figure 1, 2, 2a, 2b. Thus, we argue that the convex collection can sufficiently characterize the behavior of trained models for our purposes.

3.4 Expected Accuracy and Agreement

We are interested in two quantities: accuracy and agreement. Accuracy measures the rate at which a model's prediction agrees with the ground truth label, whereas agreement measures the rate at which the predictions of two models agree. Formally, given some distribution D over $(\boldsymbol{x}, y)$ input-label pairs and a pair of models f_i and f_j that map from $\mathbb{R}^d \to \{-1, 1\}$, the expected accuracy and agreement is defined as

$$\mathrm{Acc}_D(f_i) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\nVdash\{f_i(x) = y\}\right], \tag{3}$$

$$\mathrm{Agr}_D(f_i, f_j) = \mathbb{E}_{x,y\sim\mathcal{D}}\left[\nVdash\{f_i(x) = f_j(x)\}\right], \tag{4}$$

Over our binary Gaussian mixture, for any pair of linear models $\boldsymbol{w}_i$ and $\boldsymbol{w}_j$, these quantities simplify to

$$\mathrm{Acc}_D(\boldsymbol{w}_i) = \mathrm{Pr}_{x,y\sim D}\left(y \cdot \boldsymbol{w}_i^T\boldsymbol{x} \geq 0\right) = \Phi(x_i) \tag{5}$$

$$\begin{aligned}\mathrm{Agr}_D(\boldsymbol{w}_i, \boldsymbol{w}_j) &= \mathrm{Pr}_{x\sim D}\left((\boldsymbol{w}_i^T\boldsymbol{x}) \cdot (\boldsymbol{w}_j^T\boldsymbol{x}) \geq 0\right) \\ &= \mathrm{Pr}_{x\sim D}\left(\boldsymbol{w}_i^T\boldsymbol{x} \geq 0, \boldsymbol{w}_j^T\boldsymbol{x} \geq 0\right) \\ &\quad + \mathrm{Pr}_{x\sim D}\left(\boldsymbol{w}_i^T\boldsymbol{x} < 0, \boldsymbol{w}_j^T\boldsymbol{x} < 0\right) \\ &= \mathrm{BvN}(-x_i, -x_j; \rho) + \mathrm{BvN}(x_i, x_j; \rho)\end{aligned} \tag{6}$$

$$\text{where } x_i = \frac{\boldsymbol{w}_i^T\boldsymbol{\mu}}{\sigma\,\|\Sigma^{1/2}\boldsymbol{w}_i\|},\ x_j = \frac{\boldsymbol{w}_j^T\boldsymbol{\mu}}{\sigma\,\|\Sigma^{1/2}\boldsymbol{w}_j\|},$$

$$\rho = S_C\left(\Sigma^{1/2}\boldsymbol{w}_i, \Sigma^{1/2}\boldsymbol{w}_j\right)$$

and $S_C(\boldsymbol{u}, \boldsymbol{v}) = \langle \boldsymbol{u}/\|\boldsymbol{u}\|, \boldsymbol{v}/\|\boldsymbol{v}\| \rangle$. Notably, accuracy can be expressed as the standard univariate Gaussian CDF $\Phi(\cdot)$, whereas agreement is equal to the sum of two standard bivariate Gaussian CDF's $\mathrm{BvN}(\cdot)$:

$$\mathrm{BvN}(a, b; \rho) = \Pr\left(\mathcal{N}\left(\mathbf{0}, \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}\right) \leq \begin{bmatrix}a \\ b\end{bmatrix}\right)$$

$$= \frac{1}{2\pi\sqrt{1-\rho^2}}\int_{-\infty}^{a}\int_{-\infty}^{b}\exp\left[-\left(\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right)\right]dxdy$$

These derivation follow directly from the fact that under the class-conditioned distribution $\boldsymbol{x} \mid y$, $\boldsymbol{w}_i^T\boldsymbol{x}$ is also normally distributed $\mathcal{N}(y\boldsymbol{w}_i^T\boldsymbol{\mu}, \sigma^2\boldsymbol{w}_i^T\Sigma\boldsymbol{w}_i)$ due to the linear transformation property.

3.5 Probit Transform

To observe a better linear trend in ID versus OOD accuracy and agreement, it is common practice to first transform the statistics by the probit scale, i.e., the inverse CDF of the standard univariate Gaussian $\Phi^{-1}(\cdot)$ (Miller et al., 2021; Baek et al., 2022). In our paper, we similarly study linear trends after probit scaling.

3.6 Accuracy and Agreement on the Line

We now formally describe the two phenomena. *Accuracy-on-the-line* refers to when the the probit-scaled OOD accuracy is strongly linearly correlated with the probit-scaled ID accuracy across a set of n models. Specifically, $\forall i \in [n]$,

$$\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w}_i)\right) = \mathrm{m} \cdot \Phi^{-1}\left(\mathrm{Acc}_D(\boldsymbol{w}_i)\right) + \mathrm{b}$$

Previous empirical work fit the slope and bias via least squares over the collection of trained classifiers. In this work, we simply define the linear fit as the line connecting the ID versus OOD accuracy of the optimal Bayes classifier and random classifiers in $\mathcal{H}_{0,1}$. Then the slope and bias of accuracy-on-the-line is

$$\mathrm{Slope(m)} \coloneqq \frac{\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w}^{\star})\right) - \mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w})\right)\right]}{\Phi^{-1}\left(\mathrm{Acc}_{D}(\boldsymbol{w}^{\star})\right) - \mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D}(\boldsymbol{w})\right)\right]}$$

$$\begin{aligned}\mathrm{Bias(b)} \coloneqq\ &\mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w})\right)\right] - \\ &\mathrm{m} \cdot \mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D}(\boldsymbol{w})\right)\right]\end{aligned} \tag{7}$$

Note that the quantities $\mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w})\right)\right]$ and $\mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D}(\boldsymbol{w})\right)\right]$ over all random initializations $\boldsymbol{\epsilon} \sim \mathcal{W}$ are exactly equal to 0 in our theoretical setting. In short, this is because the expected accuracy of random classifiers is 50% and $\Phi^{-1}(0.5) = 0$. This further means that the linear trend is defined by the slope only. The slope is defined as the ratio between the performance ID and OOD of the Bayes optimal classifier. The *strength* of accuracy-on-the-line is measured by the magnitude of the absolute residual away from the line. For any classifier $\boldsymbol{w}_i \in \mathcal{H}$, the accuracy-on-the-line absolute residual is

$$\mathrm{R}_{\mathrm{Acc}}(\boldsymbol{w}_i) = \left|\Phi^{-1}\left(\mathrm{Acc}_{D'}(\boldsymbol{w}_i)\right) - \mathrm{m} \cdot \Phi^{-1}\left(\mathrm{Acc}_D(\boldsymbol{w}_i)\right)\right|$$

Agreement-on-the-line is a coupled phenomenon comprised of two parts. *First*, when accuracy-on-the-line is strong, meaning the accuracy residual is negligible, the ID versus OOD agreement must also be strongly linearly correlated with the same slope and bias. *Second*, when accuracy-on-the-line holds weakly, meaning the residual $\mathrm{R}_{\mathrm{Acc}}$ is large, the residual of ID versus OOD agreement must also be weak, such that there is no false positive event where agreement over-promises accuracy-on-the-line. To capture both of these qualities, we measure the absolute residual of ID versus OOD agreement of any pair of models $\boldsymbol{w}_i, \boldsymbol{w}_j \sim \mathcal{H}$ away from the linear fit as defined by the *slope of accuracy-on-the-line*:

$$\begin{aligned}&\mathrm{R}_{\mathrm{Agr}}(\boldsymbol{w}_i, \boldsymbol{w}_j) = \\ &\left|\Phi^{-1}\left(\mathrm{Agr}_{D'}(\boldsymbol{w}_i, \boldsymbol{w}_j)\right) - \mathrm{m} \cdot \Phi^{-1}\left(\mathrm{Agr}_D(\boldsymbol{w}_i, \boldsymbol{w}_j)\right)\right|\end{aligned}$$

We will also denote the signed residual without the absolute value as $\tilde{\mathrm{R}}_{\mathrm{Acc}}(\cdot)$, $\tilde{\mathrm{R}}_{\mathrm{Agr}}(\cdot)$.

4 AGREEMENT-ON-THE-LINE UNDER SCALE SHIFTS

In this section, we formalize the set of conditions under which the first part of agreement-on-the-line holds, i.e., strong linear correlation with matching slopes and bias, under conditions where accuracy-on-the-line observes *a perfect linear trend*. Miller et al. (2021) showed that under distribution shifts $D \to D'$ that simply change the *scale* of the mean $\boldsymbol{\mu}' = \alpha\boldsymbol{\mu}$ or covariance $\sigma' = \gamma\sigma$, the ID versus OOD accuracy of any classifier $\boldsymbol{w} \in \mathbb{R}^d$ lies exactly on the following line:

$$\Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}\right)\right) = \frac{\boldsymbol{w}^T\boldsymbol{\mu}'}{\sigma'\left\|\Sigma^{1/2}\boldsymbol{w}\right\|} = \frac{\alpha}{\gamma}\cdot\frac{\boldsymbol{w}^T\boldsymbol{\mu}^T}{\sigma\left\|\Sigma^{1/2}\boldsymbol{w}\right\|} \tag{8}$$

$$\Rightarrow \Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}\right)\right) = \frac{\alpha}{\gamma}\Phi^{-1}\left(\mathrm{Acc}_{D}\left(\boldsymbol{w}\right)\right) \tag{9}$$

We now ask: under simple scale shifts, when do we observe agreement-on-the-line with the same slope $\mathrm{m} = \alpha/\gamma$? Does agreement-on-the-line impose any additional data and model conditions?

4.1 Numerical Computation of the Agreement Residual

While the probit-scaled accuracy can be written in closed-form, such as in Equation 8, quantifying probit-scaled agreement is difficult. From Equation 5, we know that under scale shifts $\mathrm{m} = \alpha/\gamma$, agreement for any two classifiers $\boldsymbol{w}_1, \boldsymbol{w}_2$ is equal to

$$\begin{aligned}
&\mathrm{Agr}_D\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right) = b(x_1, x_2, \rho)\\
&\mathrm{Agr}_{D'}\left(\boldsymbol{w}_1, \boldsymbol{w}_2\right) = b(\mathrm{m}x_1, \mathrm{m}x_2, \rho)\\
&\text{where } b(a, b, c) = \mathrm{BvN}\left(a, b; c\right) + \mathrm{BvN}\left(-a, -b; c\right)
\end{aligned}$$

Specifically, ID and OOD agreement is a function of three variables: the *probit-scaled ID accuracies* of the two models $x_i = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}_i\right)\right))$ and $x_j = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}_j\right)\right))$, and

$$\rho = S_C\left(\Sigma^{1/2}\boldsymbol{w}_i, \Sigma^{1/2}\boldsymbol{w}_j\right) \tag{10}$$

which is the cosine similarity between the classifiers projected onto the covariance. We refer to ρ as the *similarity score* since it roughly captures how "similar" two models are to each other inside the span of the data. Note that ρ is fixed between Agr_D and $\mathrm{Agr}_{D'}$ since we set $\Sigma = \Sigma'$. To emphasize this simplification to three variables, we will often refer to ID agreement as $b(x_1, x_2, \rho) = \mathrm{Agr}_D\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right)$ and OOD agreement as $b(\mathrm{m}x_1, \mathrm{m}x_2, \rho) = \mathrm{Agr}_{D'}\left(\boldsymbol{w}_i, \boldsymbol{w}_j\right)$, where $b(\cdot)$ is the sum of two bivariate normal CDFs, one with negated limits of the other. Finally, the agreement absolute residual is equal to

$$\begin{aligned}
&\mathrm{R}_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2)\\
&= \left|\Phi^{-1}\left(b(\mathrm{m}x_1, \mathrm{m}x_2, \rho)\right) - \mathrm{m}\Phi^{-1}\left(b(x_1, x_2, \rho)\right)\right|
\end{aligned}$$

In the following subsections, we aim to characterize the region ρ values where this residual is small (i.e., agreement-on-the-line holds) for any two classifiers with ID accuracies x_1 and x_2. However, one bottleneck is that probit-scaled agreement does not have a closed-form like probit-scaled accuracy — the probit (inverse of the univariate CDF) is unable to directly simplify bivariate CDFs in a similar fashion as univariate CDFs. We instead use the SciPy module (`scipy.stats.multivariate_normal.cdf`) to calculate $\mathrm{BvN}(\cdot)$ by numerical integration (Genz, 1992). We then compose its output with the SciPy implementation of the probit (`scipy.stats.norm.ppf`).

4.2 Similarity Score Region where Agreement-on-the-Line Holds

For a scale shift $\mathrm{m} = \alpha/\gamma$, we use fine-grained grid search to find the region of triplets $(x_1, x_2, \rho) \in \mathcal{S}$ where probit-scaled ID versus OOD agreements follow a strong linear trend with the same slope as accuracies:

$$\left|\Phi^{-1}\left(b(\mathrm{m}x_1, \mathrm{m}x_2, \rho) - \mathrm{m}\Phi^{-1}\left(b(x_1, x_2, \rho)\right)\right| \leq 0.05 \tag{11}$$

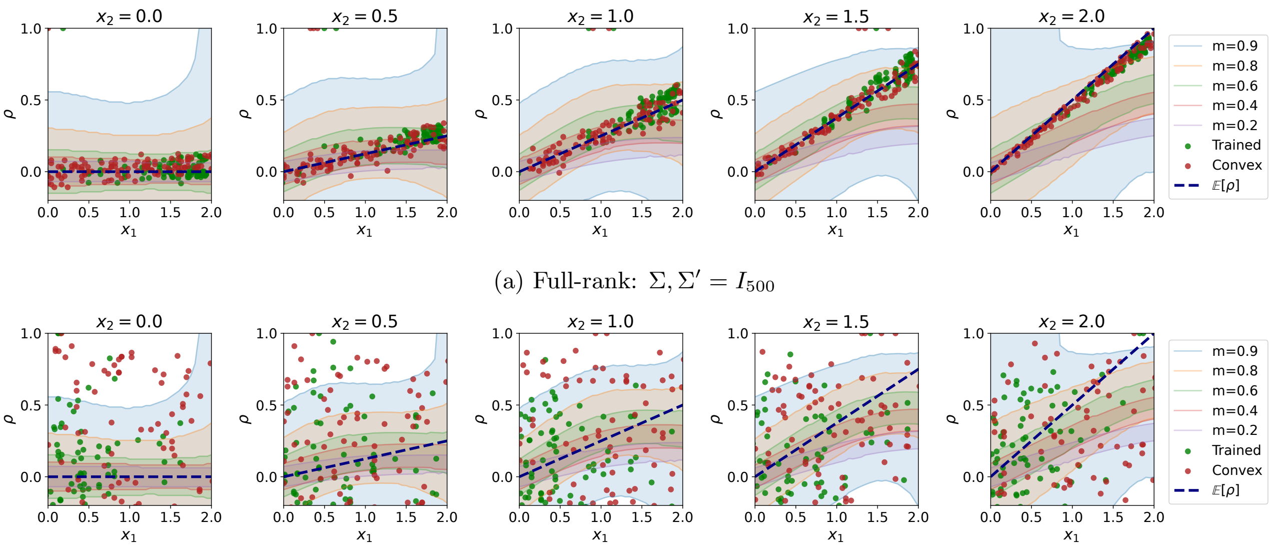
Specifically, for scale shifts $\mathrm{m} \in [0.2, 0.4, 0.6, 0.8, 0.9]$, we search over 1.) probit-scaled ID accuracies $x_1, x_2 \in [0, 2]$ and 2.) similarities $\rho \in [-1, 1]$

Our results show that under scale shifts, while accuracy-on-the-line holds perfectly for arbitrary sets of classifiers, agreement-on-the-line holds ($\mathrm{R}_{\mathrm{Agr}} < 0.05$) for a restricted set of models. In Figure 2, we fix $x_1 \in [0.0, 0.5, 1.0, 1.5, 2.0]$ and shade the region of triplets (x_1, x_2, ρ) where for a scale shift m, $\mathrm{R}_{\mathrm{Agr}} < 0.05$. We call this region of (x_1, x_2, ρ) triplets the "good region"—any set of models fall within this region would demonstrate strong accuracy and agreement-on-the-line. While for small distribution shifts ($\mathrm{m} > 0.8$), the residual is small for a fairly wide range of similarity scores ρ for any x_1, x_2, the good region *quickly narrows and moves closer to 0* as the the scale shift gets larger ($\mathrm{m} \to 0$). Indeed, in larger scale shifts, we observe that agreement-on-the-line begins to break in convex collections due to similarity scores between model pairs falling outside the good region (Figure 3). In the following conjecture, we provide a closed-form approximation of the good region to formally characterize these observations.

Numerical Result 4.1 *For any two classifiers* w_1 *and* w_2 *with probit scaled ID accuracies* $x_1, x_2 \in [0, 2]$ *respectively, under any scale shift determined by* $\mathrm{m} = \alpha/\gamma \in [0.2, 0.9]$*, if the similarity score* ρ *is within*

$$0.3x_1x_2(\sqrt{\mathrm{m}} - 0.2) \pm \delta \tag{12}$$

where $\delta = 0.1\mathrm{m} - 0.1\mathrm{m}^2(1 - x_1)(1 - x_2) + 0.5\mathrm{m}^3$*, then* $R_{Agr}(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq 0.05$.



(a) Full-rank: $\Sigma, \Sigma' = I_{500}$

(b) Approximately Low-Rank: $\Sigma, \Sigma' = \Sigma_{500}^{-2}$

Figure 2: **Good region of (x_1, x_2, ρ)** For a fixed $x_1 \in [0.0, 2.0]$, we shade the region of all (x_1, x_2, ρ) triplets where for a scale shift m, agreement-on-the-line holds with $\mathrm{R_{Agr}} < 0.05$. This region is larger for smaller scale shifts ($\mathrm{m} \geq 0.6$). In green and red points, we plot the similarity scores between pairs of trained and convex models, respectively. In dashed line, we plot the expected similarity score from Equation 16. (a) When $\Sigma = I$, similarity scores concentrate inside the good region, while (b) it varies wildly for fast-decaying covariance matrices.

Importantly, notice that the region is roughly described by the interval $f(x_1, x_2, \mathrm{m}) \pm \delta$ where f *grows bilinearly* with respect to x_1 and x_2. In §A.1, we provide a more exact non-linear closed-form approximation of $\rho^* = f(x_1, x_2)$ where $R_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2) = 0$. Moreover, we make the observation that the first derivative of f with respect to x_1 or x_2 is less than 0.3. This suggesets that agreement-on-the-line requires pairs of models have to be sufficiently uncorrelated. We will explore this further in the next section.

4.3 Convex Collection Lies Inside Good Region

From the previous section, we learned that given any two models with probit-scaled ID accuracies x_1 and x_2, there is a specific range of similarity scores where agreement-on-the-line holds, and this range narrows for larger scale shifts. Furthermore, we observed that this region is well characterized by the interval $f(x_1, x_2, \rho) \pm \delta$ where f grows bilinearly in x_1 and x_2. The bilinear approximation will become important in this section, where we analyze if models in the convex collection falls within the good region.

In Figure 2, we sample a convex collection of models with coefficients $a, b \sim \mathcal{U}[0, 1]$ and visualize the similarity score of model pairs. Surprisingly, as shown in green dots, we observe that 1.) the similarity scores tend to lie within the good region for most scale shifts, i.e., $\mathrm{m} \in \{0.4, 0.6, 0.8, 0.9\}$, when $\Sigma = I_{500}$, but 2.) when the covariance matrix is fast-decaying (e.g., $\Sigma^{500}_{5,0.01}$), we observe a higher variance in similarity scores, causing similarity scores to fall out of the good region. We corroborate our results in gradient descent trained models. Interestingly, the similarity score between trained classifiers follow the same distribution of similarity scores as our convex collection. In §A.1, we plot the good region for each value of m as separate figures for better visualization.

In the remainder of this section, we theoretically show why the distribution of similarity scores of models pairs sampled from the convex collections closely track the bilinear approximation of the "good region" $f(x_1, x_2, \rho) \pm \delta$ (Equation 4.1).

Expected similarity score We first show that the expected similarity score of the convex collection is bilinear over Gaussians with isotropic covariance matrices $\Sigma, \Sigma' = I$. Given two classifiers $w_1 \sim \mathcal{H}_{a_1,b_1}$ and $w_2 \sim \mathcal{H}_{a_2,b_2}$, their expected similarity score is

$$\rho = \left\langle \frac{\boldsymbol{w}_1}{\|\boldsymbol{w}_1\|}, \frac{\boldsymbol{w}_2}{\|\boldsymbol{w}_2\|} \right\rangle \tag{13}$$

$$= a_1 a_2 \|\boldsymbol{\mu}\|_2^{-2} \langle \boldsymbol{\mu}, \boldsymbol{\mu} \rangle + b_1 b_2 \langle \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \rangle \tag{14}$$

$$= a_1 a_2 + b_1 b_2 \langle \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \rangle \tag{15}$$

$$\Rightarrow \mathbb{E}_{w_1, w_2}[\rho] = a_1 a_2 = (\|\boldsymbol{\mu}\| / \sigma)^{-2} x_1 x_2 \tag{16}$$

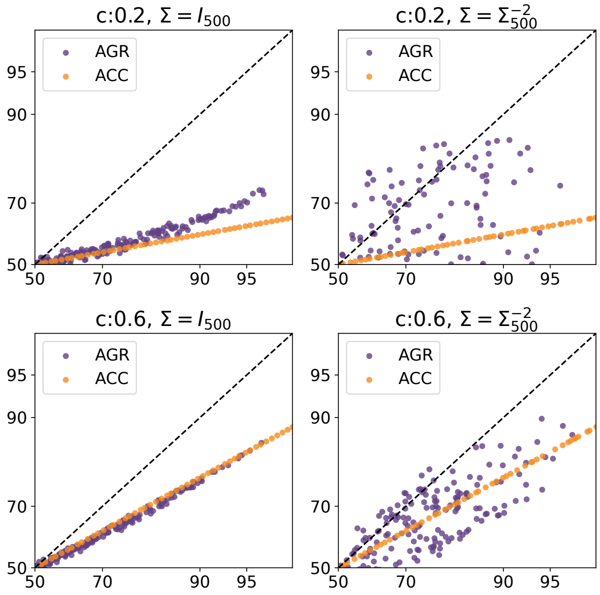


Figure 3: Agreement on the line under different levels of scale shift, where $\mathrm{m} = \sigma'/\sigma$ and covariance shapes $\Sigma = \Sigma_d^{-\alpha}$ as defined in §3.2.

Note that the $\langle \boldsymbol{\mu}, \boldsymbol{\epsilon}_i \rangle$ terms disappear since any random initialization is orthogonal to the Bayes optimal classifier by construction. Also, we rewrite the expected similarity with respect to the probit-scaled ID accuracies, which are equal to $x_i = \sigma^{-1} \|\boldsymbol{w}\|^{-1} \langle \boldsymbol{w}, \boldsymbol{\mu} \rangle = \sigma^{-1} a_i \|\boldsymbol{\mu}\|$ under $\Sigma, \Sigma' = I$. Further, note that $\|\boldsymbol{\mu}\| / \sigma$ is equal to the probit-scaled accuracy of the Bayes optimal classifier. This lends us the following result.

Proposition 4.2 *Under scale shifts* $\mathrm{m} = \alpha/\gamma < 1$ *with isotropic class-conditional Gaussians (i.e.* $\Sigma, \Sigma' = I$*), if the optimal classifier achieves*

$$\Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}^\star\right)\right) = 1.82(\sqrt{m} - 0.2)^{-1/2}$$

then for two classifiers $w_1 \sim \mathcal{H}_{a_1,b_1}$ *and* $w_2 \sim \mathcal{H}_{a_2,b_2}$, $\mathbb{E}_{w_1,w_2}[\rho]$ *is equal to the bilinear approximation of the good region* $f(x_1, x_2, \rho)$ *from Eq. 16.*

Variance of Simility Score Furthermore, we note that as the data dimension grows, the agreement between any two classifiers $w_1 \sim \mathcal{H}_{a_1,b_1}$ and $w_2 \sim \mathcal{H}_{a_2,b_2}$ concentrates at $\mathbb{E}_{w_1,w_2}[\rho]$. From Equation 15, note that the variance of ρ is

$$\mathrm{Var}_{\mathrm{w}_1,\mathrm{w}_2}(\rho) = \mathrm{b}_1^2 \mathrm{b}_2^2 \mathbb{E}\left[\langle \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \rangle^2\right]. \tag{17}$$

Using a standard concentration bound on a sphere, we show that with high probability, $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are close to orthogonal in high dimensions. Note that the set $\mathcal{W}$ where $\boldsymbol{\epsilon}$ is sampled from is simply $d-1$ dimensional sphere embedded in d dimensional space. Using the following concentration bound for a fixed unit vector $\boldsymbol{v}$: $\Pr\left(|\langle \boldsymbol{v}, \boldsymbol{\epsilon} \rangle| > z\right) \leq 2\exp\left(-(d-1)z^2/2\right)$ (Ball et al. (1997), Lemma 2.1), we prove the following proposition. See Appendix A.1 for proof.

Proposition 4.3 *Under the data assumptions in §3.2 and* $\Sigma = I$*, with probability at least* $1-\delta$*, the similarity score between classifiers* $\boldsymbol{w}_1 \sim \mathcal{H}_{a_1,b_1}$ *and* $\boldsymbol{w}_2 \sim \mathcal{H}_{a_2,b_2}$ *falls within*

$$\rho \in a_1 a_2 \pm b_1 b_2 \sqrt{2(d-1)^{-1} \log(2/\delta)} \tag{18}$$

Note that the bound grows inversely with $\sqrt{d}$, and as $d \to 1$, ρ diverges away from its expected value.

4.4 Practical Takeaways

Optimal Classifier Performance Proposition 4.2 suggests that for agreement-on-the-line to have matching slope as accuracy-on-the-line, the $\boldsymbol{w}^\star$ used to construct the convex collection must achieve high ID performance (for scale, note that $\Phi^{-1}(97.7\%) = 2$), especially for larger shifts $\mathrm{m} \ll 1$. For a more performant $\boldsymbol{w}^\star$, a model does not have to stray less away from the random initialization to achieve some $x\%$ accuracy. Thus, pairs of models with fixed ID accuracies x_1, x_2 are less similar, and we saw that low ρ is necessary for agreement-on-the-line in Equation 12. We also note that this reflects empirical findings in Figure 1 of Baek et al. (2022). In benchmarks where agreement-on-the-line appears in deep networks, the best ID model either achieves $\geq 0.9\%$ ID accuracy (e.g., CIFAR10C) or the distribution shift is small (e.g., fMoW).

Degree of Distribution Shift Agreement-on-the-line also depends on the scale of the distribution shift. As we saw from Figure 2, the good region narrows for larger shifts $\mathrm{m} \ll 1$, causing the similarity score between pairs of models to fall outside this region. This is reflected in our approximation of this region from Equation 4.1 and Figure 6, where we find see the width of this region $\delta \approx \mathcal{O}(m^3)$.

Eigenvalue Decay Rate of Covariance Matrix Proposition A.1 tells us that agreement is strongly linearly correlated in convex collections in high data dimension. This has important constraints on the shape of the covariance matrix. As shown in Figure 2, when the covariance matrix is *fast-decaying* or low rank, we similarly see a large variance in the similarity score.

5 JOINT OCCURRENCE OF ON-THE-LINE TRENDS

In this section, we prove the weak-correlation case of agreement-on-the-line. Recall that agreement-on-the-line is a two-part empirical phenomenon—either ID versus OOD accuracy and agreement are both strongly linearly correlated with matching linear fits, or they are both weakly linearly correlated with roughly the same correlation coefficient value R^2. In practice, the correlation of agreement is a useful measure for estimating if accuracy-on-the-line holds without any labels. While it is difficult to show that the R^2 of accuracy and agreement match exactly, we prove a weaker statement – for any distribution shift, the worst residual of agreement $\mathrm{R}_{\mathrm{Agr}}$ between any two classifiers is often *at least as large* as the worst residual of accuracy $\mathrm{R}_{\mathrm{Acc}}$. This guarantees that there is no false positive event where agreement-on-the-line holds but accuracy-on-the-line does not.

Consider a general distribution shift where accuracy-on-the-line does *not* hold perfectly. Such distribution shifts go beyond scale shifts and include directional changes in class-conditional means and covariances such as $\boldsymbol{\mu} = \alpha\boldsymbol{\mu}' + \beta\boldsymbol{\Delta}$ or $\Sigma \neq \Sigma'$. Under such cases, we may decompose the signed agreement residual $\tilde{\mathrm{R}}_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2)$ into the following three terms:

$$\begin{aligned}
&\tilde{\mathrm{R}}_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2) \\
&\quad = \underbrace{\mathrm{m}\Phi^{-1}\left(b(x_1, x_2, \rho)\right) - \Phi^{-1}\left(b(\mathrm{m}x_1, \mathrm{m}x_2, \rho)\right)}_{\text{scale-shift residual } s(\boldsymbol{w}_1, \boldsymbol{w}_2)} \\
&\quad + \underbrace{\Phi^{-1}\left(b(\mathrm{m}x_1, \mathrm{m}x_2, \rho)\right) - \Phi^{-1}\left(b(\mathrm{m}x_1 + \delta_1, \mathrm{m}x_2 + \delta_2, \rho)\right)}_{\text{perturbation residual } p(\boldsymbol{w}_1, \boldsymbol{w}_2)} \\
&\quad + \underbrace{\begin{aligned}&\Phi^{-1}\left(b(\mathrm{m}x_1 + \delta_1, \mathrm{m}x_2 + \delta_2, \rho)\right) \\ &\qquad - \Phi^{-1}\left(b(\mathrm{m}x_1 + \delta_1, \mathrm{m}x_2 + \delta_2, \rho')\right)\end{aligned}}_{\text{covariance-shift residual } c(\boldsymbol{w}_1, \boldsymbol{w}_2)}
\end{aligned}$$

where $\tilde{\mathrm{R}}_{\mathrm{Acc}}(\boldsymbol{w}_1) = \delta_1$, $\tilde{\mathrm{R}}_{\mathrm{Acc}}(\boldsymbol{w}_2) = \delta'$, and $\rho' = S_C\left(\Sigma'^{1/2}\boldsymbol{w}_1, \Sigma'^{1/2}\boldsymbol{w}_2\right)$. The first term is the scale shifts residual, the second term is from accuracy-on-the-line not holding perfectly, and the third term measures covariance shifts $\Sigma \neq \Sigma'$ leading to $\rho \neq \rho'$.

We care about the second term, or the perturbation residual $p(\boldsymbol{w}_1, \boldsymbol{w}_2)$, and how this function grows with the accuracy residual δ_1, δ_2. Specifically, we show that within models in a convex collection $\mathcal{M} = \{\boldsymbol{w}_i\}_{i=1}^n$, the *worst perturbation residual* across all pairs of models, i.e. $\max_{i,j} |p(\boldsymbol{w}_i, \boldsymbol{w}_j)|$, closely tracks the worst accuracy residual $\delta_{\max} = \max_i R(\boldsymbol{w})$.

5.1 Lower Bound of Perturbation Residual

Given a convex collection $\mathcal{M}$, say that some *bad* model $\boldsymbol{w}_{\mathrm{bad}} \in \mathcal{H}$ with probit scaled ID accuracy $\Phi^{-1}\left(\mathrm{Acc}_D\left(\hat{\boldsymbol{w}}\right)\right) = x$ achieves the largest accuracy residual $\delta_{\max} = \arg\max R(\boldsymbol{w})$. Furthermore, we can lower bound the largest perturbation residual with the perturbation residual between $\boldsymbol{w}_{\mathrm{bad}}$ and $\boldsymbol{w}^\star$, assuming that the convex collection samples the optimal classifier by setting $a = 1$ in Equation 2.

Theorem 5.1 *Say the largest residual for accuracy-on-the-line is achieved by a model $\boldsymbol{w}_{bad}$ in the set $\mathcal{M} = [\boldsymbol{w}_i]_{i=1}^n$. In other words, $|\delta_{bad}| = \max_i R(\boldsymbol{w}_i)$ and furthermore, $\delta_{bad} < 0$. Then the worst-case perturbation residual is lower bounded by*

$$\begin{aligned}
&\max_{i,j \in [n]} |p(\boldsymbol{w}_i, \boldsymbol{w}_j)| \;\geq \qquad\qquad (19) \\
&\left[\Phi\left(\mathrm{m}\sqrt{x_*^2 - x_{bad}^2}\right) - \Phi\left(-\mathrm{m}\sqrt{x_*^2 - x_{bad}^2}\right)\right] |\delta_{bad}|
\end{aligned}$$

where $x_* = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}^\star\right)\right),\; x_{bad} = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}_{bad}\right)\right)$.

See proof in Appendix 5. First, the theorem implies that the worst perturbation residual grows $\otimes(\delta_{bad})$. Second, when the distribution shift is small $\mathrm{m} \to 1$ and the accuracy gap between $\boldsymbol{w}^*$ and $\boldsymbol{w}_{bad}$ is large, the worst perturbation residual almost matches the worst accuracy residual. Notably, in the limit, as $x_* \to \infty$, the lower bound of the perturbation residual is exactly equal to δ_{bad}. This is trivially true since the agreement rate between predictions and ground truth labels is precisely accuracy. We also note that $\sqrt{x_*^2 - x_{\mathrm{bad}}^2}$ is often large in practice. Specifically, the state-of-the-art model is well above 90% on benchmarks while $\boldsymbol{w}_{\mathrm{bad}}$ tends to be a poor classifier.

6 EXPLAINING REAL-WORLD FAILURES

Finally, the conditions we have formalized under our Gaussian data and linear model setting is strongly tied to when agreement-on-the-line occurs in practice. Similar to Saxena et al. (2024), we train linear models over the CIFAR-10 representations from OpenCLIP ViT-B-32 (Ilharco et al., 2021; Dosovitskiy, 2020) pretrained on LAION-2B (Schuhmann et al., 2022). We observe accuracy-on-the-line and agreement-on-the-line on the CIFAR-10C benchmark consisting of 19 synthetic corruptions of CIFAR-10. While accuracy-on-the-line and agreement-on-the-line hold with strong linear correlation for most synthetic corruptions, there are 6 notable failure shifts (e.g., Gaussian Noise, Glass Blur, Shot Noise) where accuracy-on-the-line occurs but agreement has weaker linear correlation as measured by R^2.

Recall that under scale shifts, agreement-on-the-line may not hold when the covariance matrix is very fast-decaying inducing high variance in agreement. Surprisingly, when we compute the eigenvalues of the average empirical class covariance

$$\Sigma_{\text{Class-Avg}} = \frac{1}{K}\sum_{y=1}^{K}\sum_{i=1}^{n_y}\frac{1}{n_y}\left[\boldsymbol{z}_i^y(\boldsymbol{z}_i^y)^T - \boldsymbol{\mu}_y\boldsymbol{\mu}_y^T\right] \qquad (20)$$

where $\{\boldsymbol{z}_i^y\}_{i=1}^{n_y}$ are the CLIP representations of the examples in each class y and $\boldsymbol{\mu}_y$ is the empirical class mean, we can see a notable divide between how fast

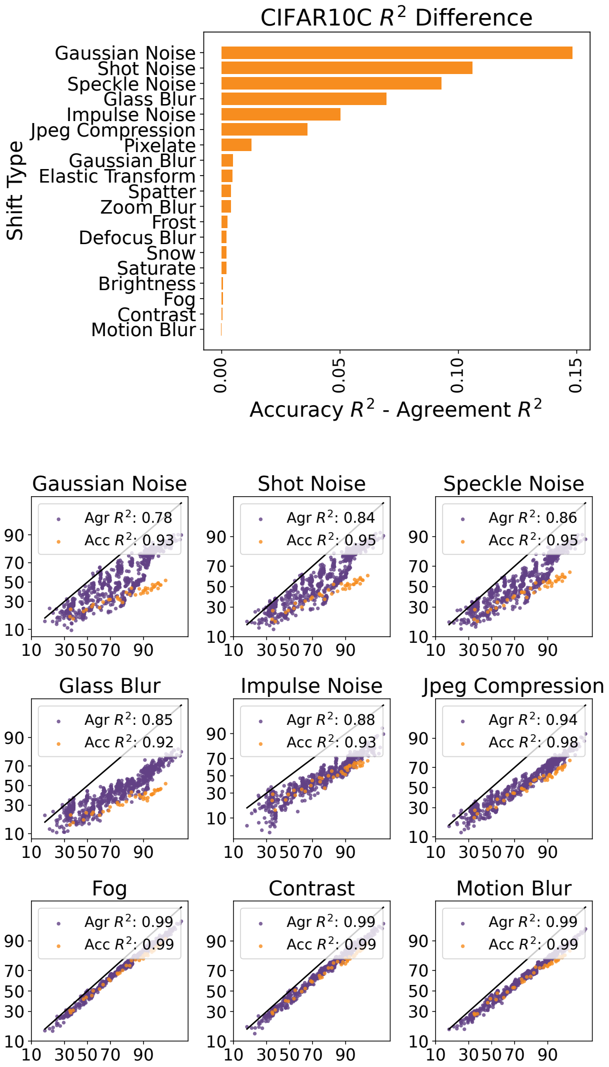


Figure 4: **CLIP Failure Cases** (Top) Difference between R^2 of ID versus OOD accuracy and agreement of CLIP linear probes for CIFAR10C shifts. (Bottom) On-the-line trends on the 6 shifts with largest R^2 difference and 3 shifts with smallest R^2 difference.

the eigenvalues decay for the failure shifts. Specifically, failure shifts have covariance matrices with faster decaying eigenvalues. Furthermore, the weakest linear correlation in agreement corresponds to the largest distribution shifts (i.e., slope of accuracy-on-the-line is small). These findings directly support our theoretical conclusions from §4.4.

7 CONCLUSION

In total, our work demonstrates that agreement-on-the-line can appear in simple linear models under the right assumptions on the randomness in the model hypotheses distribution, dimension of the data, shape of the covariance matrix, and the magnitude of the distribution shift. Our notion of having sufficiently small model similarity is closely tied to empirical observations that deep networks make uncorrelated mistakes (Nakkiran and Bansal, 2020; Jordan, 2024). In practice, the actual slope and biases of agreement and accuracy often don't match when the linear correlation is not strong, which we do not carefully formalize. We leave this for future work. Overall, we hope that our work provides new insights about how models behave under distribution shift and when we may utilize tools such as agreement-on-the-line.

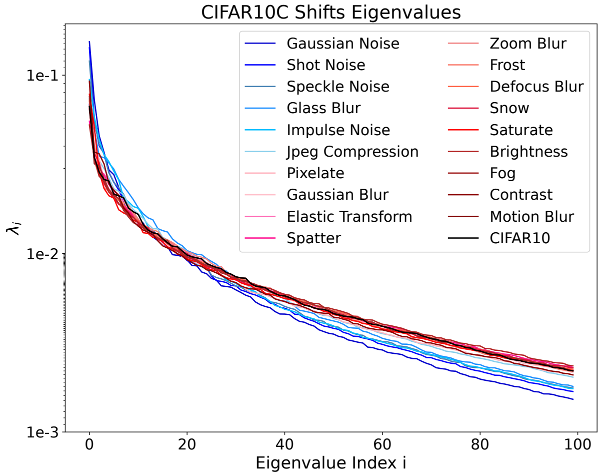


Figure 5: **Eigenvalues** Eigenvalues of $\Sigma_{\text{Class-Avg}}$ of each CIFAR10C shift sorted from largest to smallest and normalized by their sum. The eigenvalues of bad shifts (in blue) decay faster than other shifts (in red).

8 ACKNOWLEDGEMENTS

AR gratefully acknowledges support from the AI2050 program at Schmidt Sciences (Grant #G2264481), Google Research Scholar, and Apple. CB is supported by funding from the Bosch Center for Artificial Intelligence.

References

Abramowitz, M. (1974). *Handbook of Mathematical Functions, With Formulas, Graphs, and Mathematical Tables,*. Dover Publications, Inc., USA.

Awadalla, A., Wortsman, M., Ilharco, G., Min, S., Magnusson, I., Hajishirzi, H., and Schmidt, L. (2022). Exploring the landscape of distributional robustness for question answering models. *arXiv preprint arXiv:2210.12517.*

Baek, C., Jiang, Y., Raghunathan, A., and Kolter,

Z. (2022). Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. *arXiv preprint arXiv:2206.13089*.

Ball, K. et al. (1997). An elementary introduction to modern convex geometry. *Flavors of geometry*, 31(1-58):26.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Drezner, Z. and Wesolowsky, G. O. (1990). On the computation of the bivariate normal integral. *Journal of Statistical Computation and Simulation*, 35(1-2):101–107.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., and Schmidt, L. (2021). Openclip. If you use this software, please cite it as below.

Jordan, K. (2024). On the variance of neural network training with respect to test sets and distributions. In *The Twelfth International Conference on Learning Representations*.

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

Lee, D., Moniri, B., Huang, X., Dobriban, E., and Hassani, H. (2023). Demystifying disagreement-on-the-line in high dimensions.

LeJeune, D., Liu, J., and Heckel, R. (2024). Monotonic risk relationships under distribution shifts for regularized risk minimization. *Journal of Machine Learning Research*, 25(54):1–37.

Mania, H. and Sra, S. (2020). Why do classifier accuracies show linear trends under distribution shift? *CoRR*, abs/2012.15483.

Miller, J., Krauth, K., Recht, B., and Schmidt, L. (2020). The effect of natural distribution shift on question answering models. In *International conference on machine learning*, pages 6905–6916. PMLR.

Miller, J., Taori, R., Raghunathan, A., Sagawa, S., Koh, P. W., Shankar, V., Liang, P., Carmon, Y., and Schmidt, L. (2021). Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization. *International Conference on Machine Learning*.

Nakkiran, P. and Bansal, Y. (2020). Distributional generalization: A new kind of generalization. *arXiv preprint arXiv:2009.08092*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., and Schmidt, L. (2019). A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems*, 32.

Saxena, R., Kim, T., Mehra, A., Baek, C., Kolter, Z., and Raghunathan, A. (2024). Predicting the performance of foundation models via agreement-on-the-line.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C. W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S. R., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. (2022). LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shankar, V., Roelofs, R., Mania, H., Fang, A., Recht, B., and Schmidt, L. (2020). Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*, pages 8634–8644. PMLR.

Willink, R. (2005). Bounds on the bivariate normal distribution function. *Communications in Statistics-Theory and Methods*, 33(10):2281–2297.

Yadav, C. and Bottou, L. (2019). Cold case: The lost mnist digits. *Advances in neural information processing systems*, 32.

A APPENDIX

A.1 SCALE SHIFTS EXTENDED

A.1.1 VARIANCE OF SIMILARITY SCORES

Proposition A.1 *Under the data assumptions in §3.2 and* $\Sigma = I$*, with probability at least* $1 - \delta$*, the similarity score between classifiers* $\boldsymbol{w}_1 \sim \mathcal{H}_{a_1,b_1}$ *and* $\boldsymbol{w}_2 \sim \mathcal{H}_{a_2,b_2}$ *falls within*

$$\rho \in a_1 a_2 \pm b_1 b_2 \sqrt{2(d-1)^{-1} \log(2/\delta)} \tag{21}$$

Proof Given $\boldsymbol{w}_1 \sim \mathcal{H}_{a_1,b_1}$ and $\boldsymbol{w}_2 \sim \mathcal{H}_{a_2,b_2}$, recall from Equation 17 that

$$\mathrm{Var}_{\mathrm{w}_1,\mathrm{w}_2}(\rho) = \mathrm{b}_1^2 \mathrm{b}_2^2 \mathbb{E}\left[\langle \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \rangle^2\right]. \tag{22}$$

Using a standard concentration bound on a sphere, we show that with high probability, $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are close to orthogonal in high dimensions. Note that the set $\mathcal{W}$ where $\boldsymbol{\epsilon}$ is sampled from is simply $d-1$ dimensional sphere embedded in d dimensional space. Using the following concentration bound for a fixed unit vector $\boldsymbol{v}$: $\Pr\left(|\langle \boldsymbol{v}, \boldsymbol{\epsilon} \rangle| > z\right) \leq 2\exp\left(-(d-1)z^2/2\right)$ (Ball et al. (1997), Lemma 2.1), note that with probability δ

$$\Pr\left(|\langle \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2 \rangle| > z\right) \leq 2\exp\left(-(d-1)z^2/2\right) = \delta \tag{23}$$

$$\Rightarrow z = \sqrt{2(d-1)^{-1} \log(2/\delta)} \tag{24}$$

So with probability at least $1 - \delta$, ρ is within the region in Proposition A.1.

A.1.2 CLOSER APPROXIMATION OF SIMILARITY SCORES

In §4, given the ID accuracy of any two classifiers $\mathrm{Acc}_D(\boldsymbol{w}_1) = a, \mathrm{Acc}_D(\boldsymbol{w}_2) = b$ and their model similarity ρ, we used numerical estimation to identify the set of all (a, b, ρ) pairs where the residual of agreement $R_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq 0.05$. Through numerical simulations, for each degree of scale shift $c = \alpha/\gamma$, we identify a unique $\rho^* \in [-1, 1]$ for each (a, b) pair where $R_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2) = 0$. Within the range $a, b \in [0, 2]$, we see that $\rho^* = f(a, b, c)$ behaves as a smooth continuous function. We provide an approximation of f in the following fact we have empirically verified by grid search. In Fig. 6, we can see that our estimate $\tilde{f}$ tracks f closely for scale shifts $c \in [0.1, 0.9]$

Numerical Result A.2 *For any two classifiers* w_1 *and* w_2 *with probit scaled in-distribution accuracies* $a, b \in [0, 2]$ *respectively, under any scale shift determined by* $c = \alpha/\gamma \in [0.2, 0.9]$*, if the covariance-projected correlation* ρ *is within*

$$\tilde{f}(a, b, c) \pm [0.04 + 0.3(c - 0.15)^3]$$

where $\tilde{f}(a, b, c)$ *equals*

$$0.61 \log(c + 0.95)^{0.64} \exp\left(-2\sqrt{c} \log(a+1) \log(b+1) \log\left(\frac{a+1}{b+1}\right)^2\right) \left[\log(1 + a^3) \log(1 + a^3)\right]^{0.45 - 0.22\sqrt{c - 0.1}}, \tag{25}$$

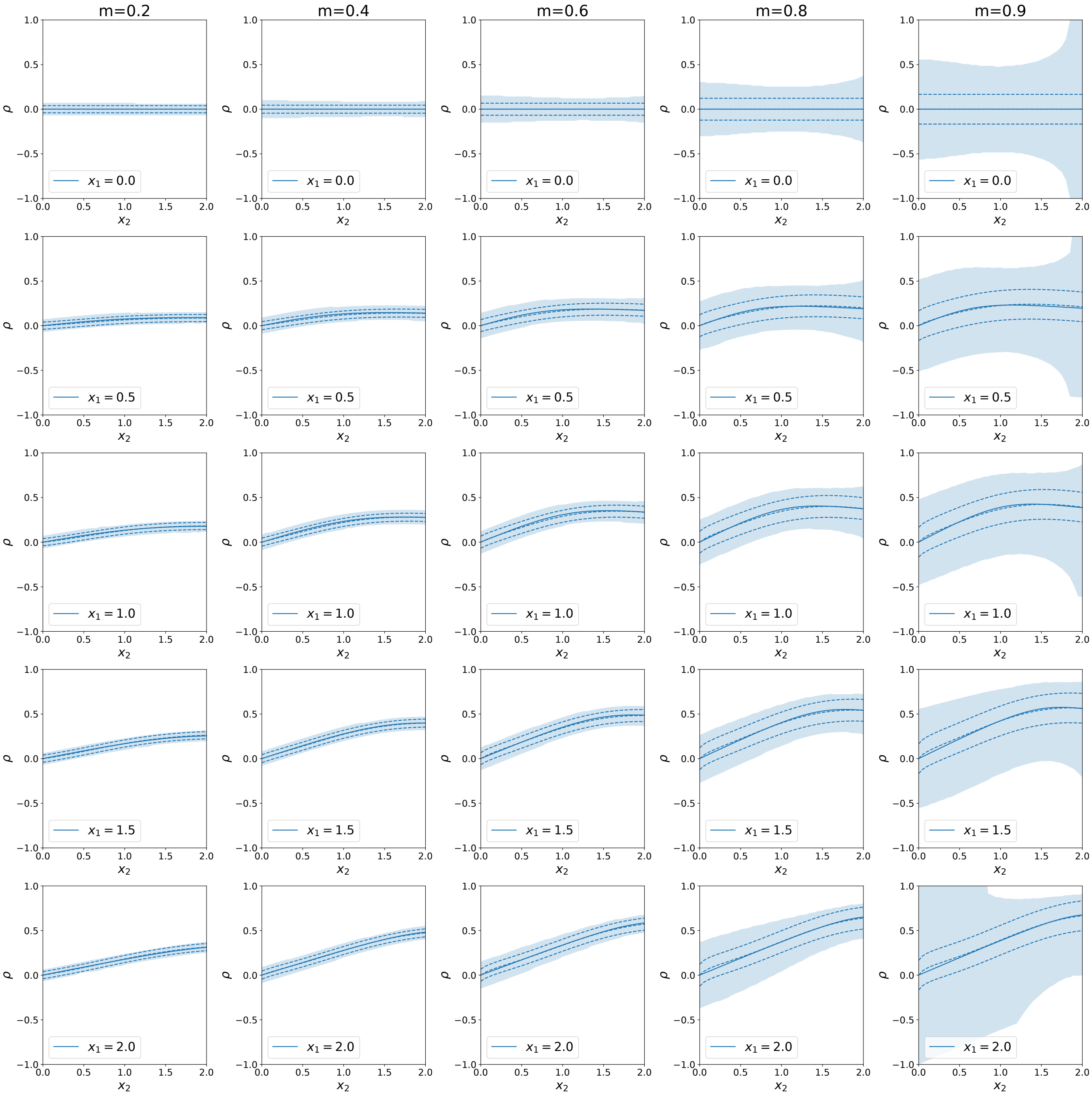


Figure 6: For each scale shift c and fixed ID accuracy of the first classifier $\mathrm{Acc}_D\left(\boldsymbol{w}_1\right) = a$, we interpolate over $\mathrm{Acc}_D\left(\boldsymbol{w}_2\right) = b \in [0, 2]$ and $\rho \in [-1, 1]$ to find all (a, b, ρ) tuples where $R_{\mathrm{Agr}}(\boldsymbol{w}_1, \boldsymbol{w}_2) \leq 0.05$. This set is shaded in light blue. In solid blue, we plot $\rho^* = f(c, a, b)$. And in dashed blue lines we plot our estimate $\tilde{f}(c, a, b)$ plus or minus the confidence interval in Fact A.1.

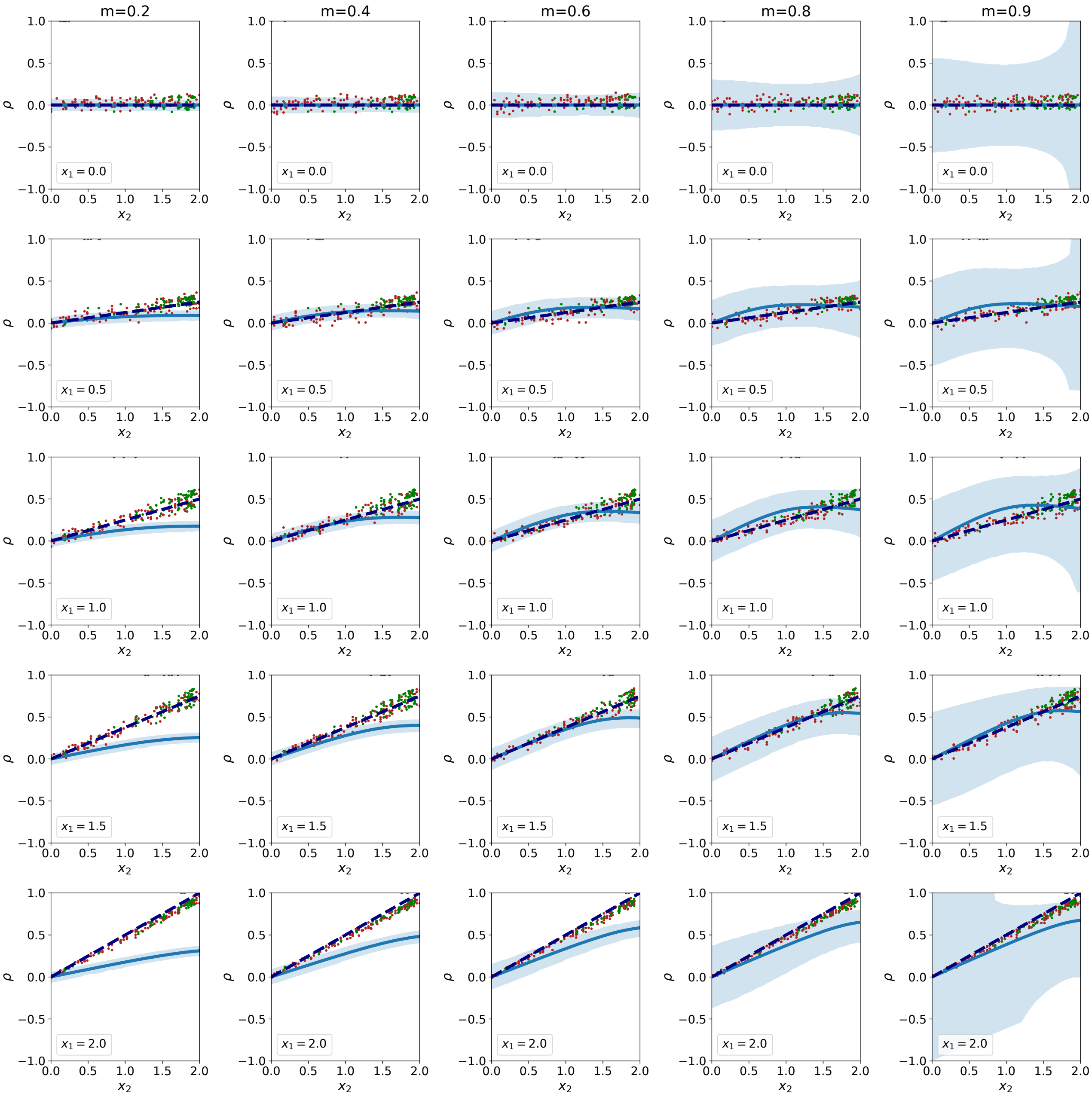


Figure 7: Full-rank: $\Sigma = I_{500}$. We provided an expanded view of Figure 2a, where we plot the good region for each degree of scale shift c in a separate subfigure. In pink points, we plot the model similarity between sampled pairs of trained classifiers optimized by gradient descent over the ID distribution. In green points, we plot the model similarity between sampled pairs from the convex ensemble. Note that the model similarities for pairs of trained classifiers and convex classifiers are similarly distributed, and they are concentrated around the expected model similarity plotted in red dashed.

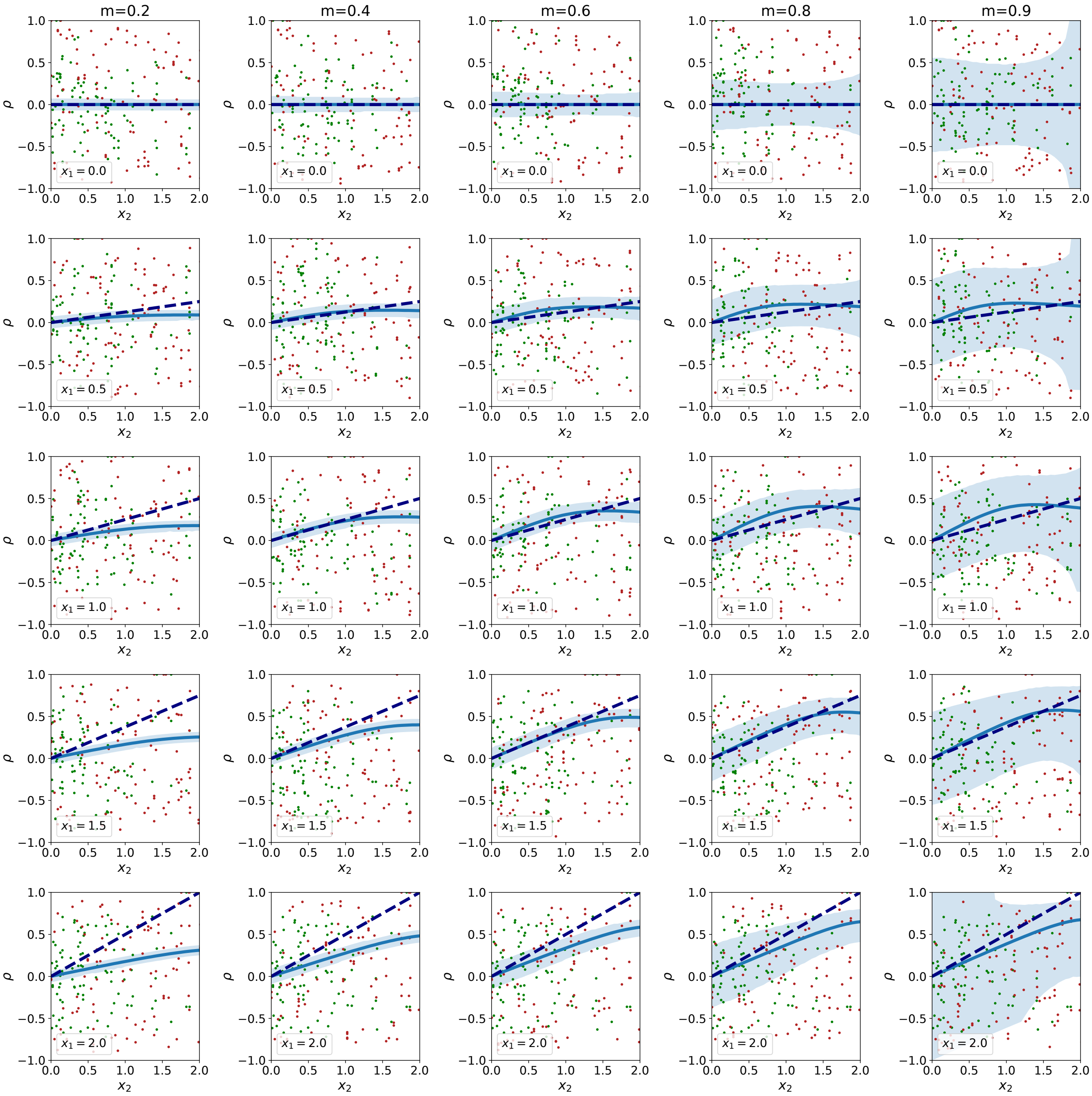


Figure 8: Approximately Low-Rank: $\Sigma = \Sigma_{500}^{-2}$ We provided an expanded view of Figure 2b, where we plot the good region for each degree of scale shift c in a separate subfigure. In pink points, we plot the model similarity between sampled pairs of trained classifiers optimized by gradient descent over the ID distribution. In green points, we plot the model similarity between sampled pairs from the convex ensemble. Note that the model similarities for pairs of trained classifiers and convex classifiers are similarly distributed, and they vary widely from the expected model similarity plotted in red dashed.

A.2 PROOF OF 5.1: LOWER BOUND OF WORST-CASE PERTURBATION RESIDUAL

Note that we can lower bound the magnitude of the perturbation residual by

$$p(\boldsymbol{w}_1, \boldsymbol{w}_2) = \left|\Phi^{-1}\left(b(cx_1, cx_2, \rho)\right) - \Phi^{-1}\left(b(cx_1+\delta, cx_2+\delta', \rho)\right)\right| \geq \min_{0\leq t\leq 1}\left|\frac{d}{dt}\Phi^{-1}\left(b(cx_1+t\delta, cx_2+t\delta', \rho)\right)\right| \tag{26}$$

where $b(x_1, x_2, \rho) = \mathrm{BvN}\left(x_1, x_2; \rho\right) + \mathrm{BvN}\left(-x_1, -x_2; \rho\right)$, x_1, x_2 are the ID accuracies of $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, respectively, and ρ is the model similarity.

The derivative of $\Phi^{-1}\left(\mathrm{BvN}\left(cx_1+t\delta, cx_2+t\delta'; \rho\right) + \mathrm{BvN}\left(-cx_1-t\delta, -cx_2-t\delta'; \rho\right)\right)$ with respect to t is

$$\frac{d}{dt}\Phi^{-1}\left(\underbrace{\mathrm{BvN}\left(cx_1+t\delta, cx_2+t\delta'; \rho\right) + \mathrm{BvN}\left(-cx_1-t\delta, -cx_2-t\delta'; \rho\right)}_{z}\right) = \tag{27}$$

$$\frac{1}{\phi\left(\Phi^{-1}\left(z\right)\right)}\left[\phi(cx_1+t\delta)\Phi\left(\frac{cx_2+t\delta'-\rho(cx_1+t\delta)}{\sqrt{1-\rho^2}}\right) - \phi(-cx_1-t\delta)\Phi\left(\frac{-cx_2-t\delta'+\rho(cx_1+t\delta)}{\sqrt{1-\rho^2}}\right)\right]\delta \tag{28}$$

$$+\frac{1}{\phi\left(\Phi^{-1}\left(z\right)\right)}\left[\phi(cx_2+t\delta')\Phi\left(\frac{cx_1+t\delta-\rho(cx_2+t\delta')}{\sqrt{1-\rho^2}}\right) - \phi(-cx_2-t\delta')\Phi\left(\frac{-cx_1-t\delta+\rho(cx_2+t\delta')}{\sqrt{1-\rho^2}}\right)\right]\delta' \tag{29}$$

We apply the following formulas above:

- **Derivative of Inverse Functions** Given an invertible function $f(x)$, the derivative of its inverse function $f^{-1}(x)$ evaluated at $x = a$ is

$$[f^{-1}]'(a) = \frac{1}{f'[f^{-1}(a)]} \tag{30}$$

- **Derivative of** $\mathrm{BvN}\left(x, y; \rho\right)$ **(Drezner and Wesolowsky, 1990)** Let ϕ be the standard normal PDF, and Φ be the standard normal CDF. Then

$$\frac{\partial}{\partial x}\mathrm{BvN}\left(x, y; \rho\right) = \phi(x)\Phi\left(\frac{y-\rho x}{\sqrt{1-\rho^2}}\right), \quad \frac{\partial}{\partial y}\mathrm{BvN}\left(x, y; \rho\right) = \phi(y)\Phi\left(\frac{x-\rho y}{\sqrt{1-\rho^2}}\right) \tag{31}$$

The largest perturbation residual between any two classifiers in the convex ensemble is lower bounded by the perturbation residual between the "worst" model $\boldsymbol{w}_{\text{bad}}$ as we define in §5 and the optimal Bayes classifier $\boldsymbol{w}^*$ which is the best classifier in our convex ensemble.

Setting $\boldsymbol{w}_1 = \boldsymbol{w}_{\text{bad}}$ and $\boldsymbol{w}_2 = \boldsymbol{w}^* = \Sigma^{-1}\boldsymbol{\mu}/\left\|\Sigma^{-1}\boldsymbol{\mu}\right\|$, note that in Eq. 27, the variables then equal

$$x_1 = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}_{\text{bad}}\right)\right), \quad x_2 = \Phi^{-1}\left(\mathrm{Acc}_D\left(\boldsymbol{w}^*\right)\right) = \frac{1}{\sigma} \tag{32}$$

$$\delta = R_{\text{Acc}}(\boldsymbol{w}_{\text{bad}}) = \delta_{max}, \quad \delta' = R_{\text{Acc}}(\boldsymbol{w}^*) = 0 \tag{33}$$

$$\rho = \frac{\boldsymbol{w}^*\Sigma\boldsymbol{w}_{\text{bad}}}{\|\Sigma^{1/2}\boldsymbol{w}^*\|\,\|\Sigma^{1/2}\boldsymbol{w}_{\text{bad}}\|} = \frac{\boldsymbol{\mu}^\top\boldsymbol{w}_{\text{bad}}}{\|\Sigma^{1/2}\boldsymbol{w}_{\text{bad}}\|} = \sigma x_1 = x_1/x_2 \tag{34}$$

Note that $\delta' = 0$ by construction. In Eq. 7, we had set the slope of the accuracy linear trend to be

$$\frac{\Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}^\star\right)\right) - \mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}\right)\right)\right]}{\Phi^{-1}\left(\mathrm{Acc}_{D}\left(\boldsymbol{w}^\star\right)\right) - \mathbb{E}_{\mathcal{H}_{0,1}}\left[\Phi^{-1}\left(\mathrm{Acc}_{D}\left(\boldsymbol{w}\right)\right)\right]} = \frac{\Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}^\star\right)\right)}{\Phi^{-1}\left(\mathrm{Acc}_{D}\left(\boldsymbol{w}^\star\right)\right)} \tag{35}$$

Thus $R_{\text{Acc}}(\boldsymbol{w}^*) = \left|\Phi^{-1}\left(\mathrm{Acc}_{D'}\left(\boldsymbol{w}^*\right)\right) - \mathrm{Slope}\Phi^{-1}\left(\mathrm{Acc}_{D}\left(\boldsymbol{w}^*\right)\right)\right|$.

By substitution, we can simplify Equation 27 as

$$= \frac{\phi(cx_1+t\delta_{max})}{\phi\left(\Phi^{-1}\left(z\right)\right)}\left[\Phi\left(\frac{c(x_2^2-x_1^2)-t\delta_{max}x_1}{\sqrt{x_2^2-x_1^2}}\right) - \Phi\left(-\frac{c(x_2^2-x_1^2)-t\delta_{max}x_1}{\sqrt{x_2^2-x_1^2}}\right)\right]\delta_{max} \tag{36}$$

We first upper bound the derivative $\phi(\Phi^{-1}\left(z\right))$. Since ϕ is unimodal centered at 0 and Φ^{-1} is monotonically increasing,

$$\phi(\Phi^{-1}\left(z\right)) \leq \phi(\max\{0, \Phi^{-1}\left(z\right)\}) \leq \phi(\max\{0, \Phi^{-1}\left(\text{lower bound of } z\right)\}) \tag{37}$$

Lemma A.3 *The following lower bound of z holds when $\boldsymbol{w}_1 = \boldsymbol{w}_{bad}$ and $\boldsymbol{w}_2 = \boldsymbol{w}^*$.*

$$z = b(cx_1 + t\delta, cx_2 + t\delta', \rho) \geq \Phi\left(cx_1 + t\delta_{max}\right) - \Phi\left(-cx_2\right)\left(1 - 2\Phi\left(-\frac{t\sqrt{x_2}\delta_{max}}{\sqrt{x_2^2 - x_1^2}}\right)\right) \tag{38}$$

Proof We know the following relation about the bivariate normal CDF (Abramowitz (1974), Sec. 26.3)

$$\mathrm{BvN}\left(h, k; \rho\right) = \Phi(h) - \Phi(-k) + \mathrm{BvN}\left(-h, -k; \rho\right) \tag{39}$$

Thus,

$$b(cx_1 + t\delta_{max}, cx_2, \rho) = \Phi(cx_1 + t\delta_{max}) - \Phi(-cx_2) + 2\mathrm{BvN}\left(-cx_1 - t\delta_{max}, -cx_2; \rho\right) \tag{40}$$

Using the following lower bound from Willink (2005),

$$\mathrm{BvN}\left(-h, -k; \rho\right) \geq \Phi(-k)\Phi\left(\frac{\rho k - h}{\sqrt{1-\rho^2}}\right), \quad k > 0, \rho \geq 0 \tag{41}$$

and setting $h = cx_1 + t\delta_{max}$, $k = cx_2 = c/\sigma > 0$, and $\rho = x_1/x_2 \geq 0$, presuming $x_1 > 0$ meaning the classifier $\boldsymbol{w}_{\text{bad}}$ does not have worse than random performance in-distribution, we get

$$\mathrm{BvN}\left(-cx_1 - t\delta_{max}, -cx_2; \rho\right) \geq \Phi(-cx_2)\Phi\left(-\frac{t\delta_{max}x_2}{\sqrt{x_2^2 - x_1^2}}\right) \tag{42}$$

$$\Rightarrow b(cx_1 + t\delta_{max}, cx_2, \rho) \geq \Phi(cx_1 + t\delta_{max}) - \Phi(-cx_2)\left(1 - 2\Phi\left(-\frac{t\delta_{max}x_2}{\sqrt{x_2^2 - x_1^2}}\right)\right) \tag{43}$$

■

Using Lemma A.3, note that

$$\frac{\phi(cx_1 + t\delta_{max})}{\phi\left(\Phi^{-1}\left(z\right)\right)} \geq \frac{\phi(cx_1 + t\delta_{max})}{\phi\left(\max\left\{0, \Phi^{-1}\left(\Phi\left(cx_1 + t\delta_{max}\right) - \Phi\left(-cx_2\right)\left(1 - 2\Phi\left(-\frac{tx_2\delta_{max}}{\sqrt{x_2^2 - x_1^2}}\right)\right)\right)\right\}\right)} \tag{44}$$

When $\delta_{max} < 0$, note that $\forall t \in [0, 1]$, $1 - \Phi\left(-\frac{tx_2\delta_{max}}{\sqrt{x_2^2 - x_1^2}}\right) \leq 0$, so $\frac{\phi(cx_1 + t\delta_{max})}{\phi(\Phi^{-1}(z))} \geq 1$. Thus, we may further lower bound Equation 27 as

$$\frac{\phi(cx_1 + t\delta_{max})}{\phi\left(\Phi^{-1}\left(z\right)\right)}\left[\Phi\left(\frac{c(x_2^2 - x_1^2) - t\delta_{max}x_1}{\sqrt{x_2^2 - x_1^2}}\right) - \Phi\left(-\frac{c(x_2^2 - x_1^2) - t\delta_{max}x_1}{\sqrt{x_2^2 - x_1^2}}\right)\right]\delta_{max} \tag{45}$$

$$\geq \left[\Phi\left(c\sqrt{x_2^2 - x_1^2}\right) - \Phi\left(-c\sqrt{x_2^2 - x_1^2}\right)\right]\delta_{max} \tag{46}$$

This completes our proof.