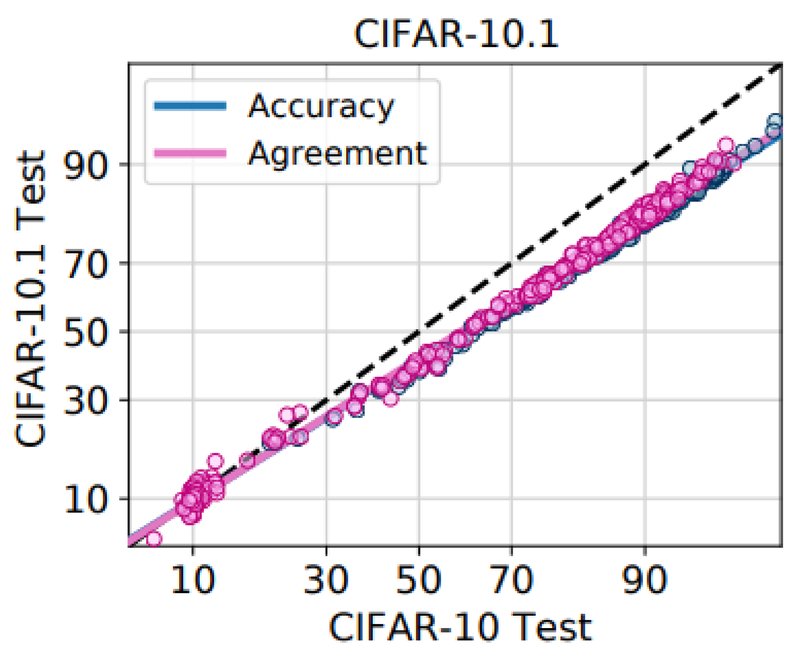
Demystifying Disagreement-on-the-Line in High Dimensions

Donghwan Lee*, Behrad Moniri*, Xinmeng Huang, Edgar Dobriban, Hamed Hassani

University of Pennsylvania

Models under Distribution Shifts

Recently, a linear trend between the ID and OOD accuracy of models has been observed. (Baek et al, 2022) found that OOD vs. ID *agreement* also forms a line, and it matches that of the accuracy.



In this paper, we study this phenomenon under a simple theoretical setting.

Theoretical Setting

Data generation: We assume that $\beta \sim \mathcal{N}(0, I_d)$

$$x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_s),\ y_i = \frac{1}{\sqrt{d}}\beta^\top x_i + \epsilon_i,\ \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$$

The input distribution *shifts* to the target distribution $x \sim \mathcal{N}(0, \Sigma_t)$ at test time.

Random features model: Two-layer neural networks with fixed, randomly generated weights in the first layer $f_{W,a}(x) = \frac{1}{\sqrt{N}} a^\top \sigma\left(Wx/\sqrt{d}\right)$.

Ridge regression: Parameters $a \in \mathbb{R}^N$ are fit via ridge regression with data $X = (x_1, \dots, x_n) \in \mathbb{R}^{d \times n}$ and $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$.

Conditions

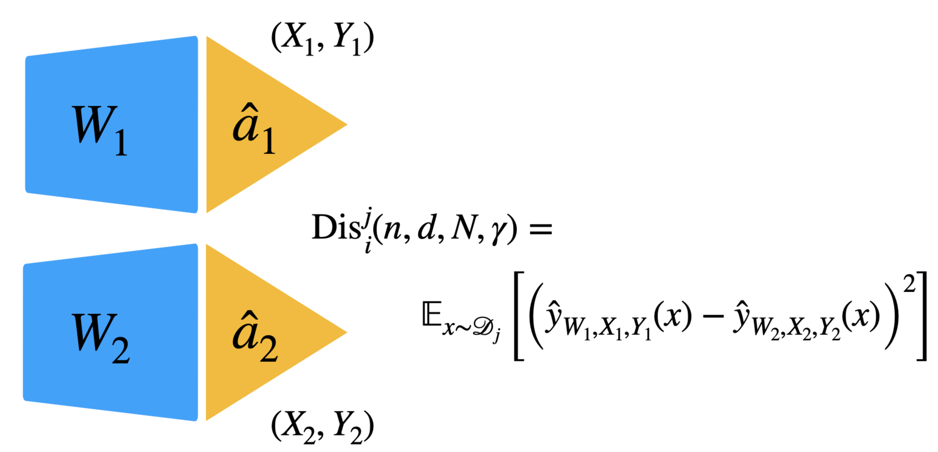
Proportional limit: We assume that $n, d, N \to \infty$ with $d/n \to \phi > 0$ and $d/N \to \psi > 0$.

Spectral property: $\Sigma_s \rightsquigarrow (\lambda_1^s, v_1), \dots, (\lambda_d^s, v_d)$. Define $\lambda_i^t = v_i^\top \Sigma_t v_i$ for $i \in [d]$. We assume that

$$\frac{1}{d}\sum_{i=1}^{d} \delta_{(\lambda_i^s, \lambda_i^t)} \to \mu.$$

Definition of Disagreement

We define disagreement as



where $j \in \{\mathrm{s}, \mathrm{t}\}$, and the index $i \in \{\mathrm{I}, \mathrm{SS}, \mathrm{SW}\}$ corresponds to one of the following cases:

- Independent disagreement ($i = \mathrm{I}$): $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$ and $W_1 \perp\!\!\!\perp W_2$.
- Shared-Sample disagreement ($i = \mathrm{SS}$): $(X_1, Y_1) = (X_2, Y_2)$ and $W_1 \perp\!\!\!\perp W_2$.
- Shared-Weight disagreement ($i = \mathrm{SW}$): $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$ and $W_1 = W_2$.

Self-Consistent Equations

Our results depend on a scalar κ, which is the solution to the *self-consistent equation*

$$\kappa = \frac{\psi + \phi - \sqrt{(\psi - \phi)^2 + 4\kappa\psi\phi\gamma/\rho_s}}{2\psi(\omega_s + \mathcal{I}_{1,1}^s(\kappa))},$$

where ρ_s and ω_s are constants depnding on the activation function, and $\mathcal{I}_{a,b}^j$ is defined by

$$\mathcal{I}_{a,b}^j(\kappa) = \phi\mathbb{E}_\mu\left[\frac{(\lambda^s)^{a-1}\lambda^j}{(\phi + \kappa\lambda^s)^b}\right], \quad j \in \{\mathrm{s}, \mathrm{t}\}.$$

Asymptotics of Disagreement

Theorem. *For the three forms of disagreement $i \in \{\mathrm{I}, \mathrm{SS}, \mathrm{SW}\}$, and $j \in \{\mathrm{s}, \mathrm{t}\}$, we provide exact asymptotic formulae for the disagreement*

$$\mathrm{Dis}_i^j(\phi, \psi, \gamma) = \lim_{n,d,N\to\infty} \mathrm{Dis}_i^j(n, d, N, \gamma),$$

We obtain a simpler expression by taking the ridgeless limit $\gamma \to 0$. The self-consistent equation simplifies to

$$\kappa = \frac{\min(1, \phi/\psi)}{\omega_s + \mathcal{I}_{1,1}^s(\kappa)}.$$

[See Theorem 3.1 and Corollary 3.2 for details]

Disagreement-on-the-Line

Recently, (Tripuraneni et al., 2021) proved that under **covariate shift**, in the **ridgeless**, and **overparameterized** regime $\phi > \psi$ we have

$$\lim_{\gamma\to 0} \mathrm{Risk}^t(\phi, \psi, \gamma) = a \lim_{\gamma\to 0} \mathrm{Risk}^s(\phi, \psi, \gamma) + b_{\mathrm{risk}},$$

where a and b_{risk} are independent of ψ.

Theorem. *In the **ridgeless**, and **overparameterized** regime $\phi > \psi$ and for $i \in \{\mathrm{I}, \mathrm{SS}\}$,*

$$\lim_{\gamma\to 0} \mathrm{Dis}_i^t(\phi, \psi, \gamma) = a \lim_{\gamma\to 0} \mathrm{Dis}_i^s(\phi, \psi, \gamma) + b_i,$$

where the slopes and intercept are independent of ψ.

[See Theorem 4.1 for details]

Approximate Linear Relation

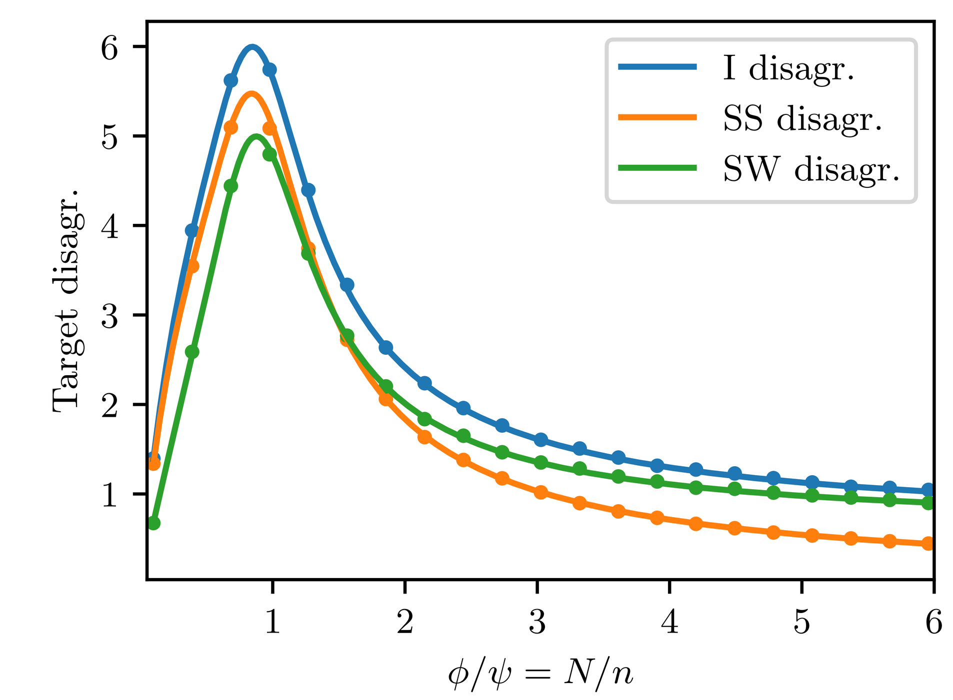
Theorem (Approximate linear relation). *Given $\phi > \psi$, deviation from the line, for I and SS disagreement, is bounded by*

$$|\mathrm{Dis}_\mathrm{I}^t(\phi, \psi, \gamma) - a\mathrm{Dis}_\mathrm{I}^s(\phi, \psi, \gamma) - b_\mathrm{I}| \leq \frac{C(\gamma + \sqrt{\psi\gamma} + \psi\gamma + \gamma\sqrt{\psi\gamma})}{(1 - \psi/\phi + \sqrt{\psi\gamma})^2},$$

$$|\mathrm{Dis}_\mathrm{SS}^t(\phi, \psi, \gamma) - a\mathrm{Dis}_\mathrm{SS}^s(\phi, \psi, \gamma)| \leq \frac{C(\sqrt{\psi\gamma} + \psi\gamma + \gamma\sqrt{\psi\gamma})}{(1 - \psi/\phi + \sqrt{\psi\gamma})^2}$$

where $C > 0$ depends on $\phi, \mu, \sigma_\epsilon^2$, and σ.

Asymptotics vs. Simulations

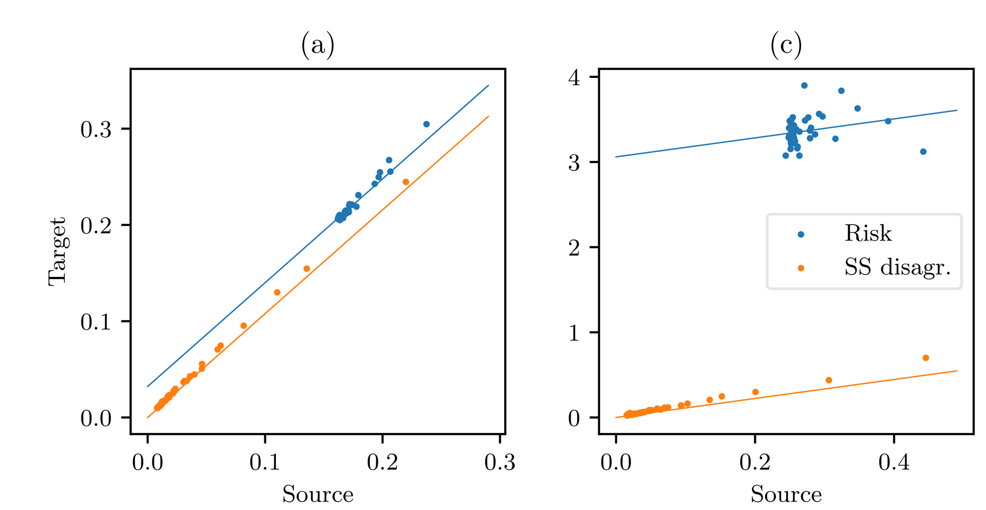


$[\gamma = 0.01, \sigma_\epsilon^2 = 0.25$, and $\mu = 0.5\delta_{(1.5,5)} + 0.5\delta_{(1,1)}, \phi = 0.5]$

Real World Experiments

Similar results hold for real-world datasets where the Gaussianity and linearity are violated.

Experiments of (a) CIFAR and (c) Camelyon17



Disagreement-on-the-Line in different regimes

(a) **underparameterized models** (b) **ridge regularization** (c) **different intercepts**