ML+FM Seminar

Demystifying *Disagreement-on-the-Line* in High Dimensions

(Joint work with Donghwan Lee, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani)

Behrad Moniri
University of Pennsylvania

April 21, 2023

Introduction

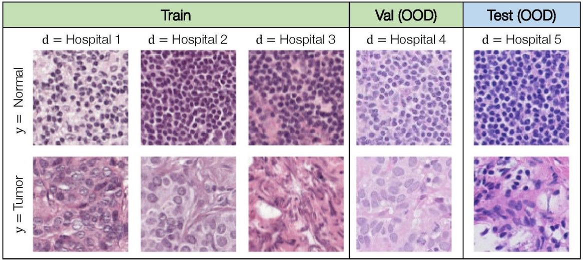
Accuracy of ML models can degrade under distribution shifts.

[Koh et al., 2021]

Introduction

Accuracy of ML models can degrade under distribution shifts.

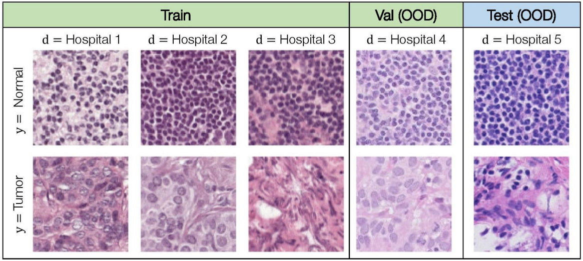
[Koh et al., 2021]

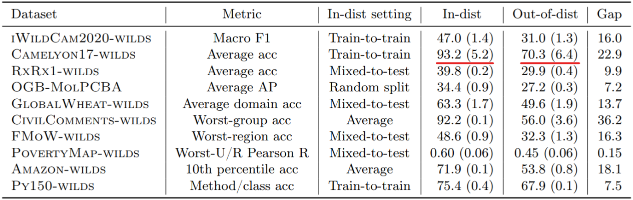


Introduction

Accuracy of ML models can degrade under distribution shifts.

[Koh et al., 2021]

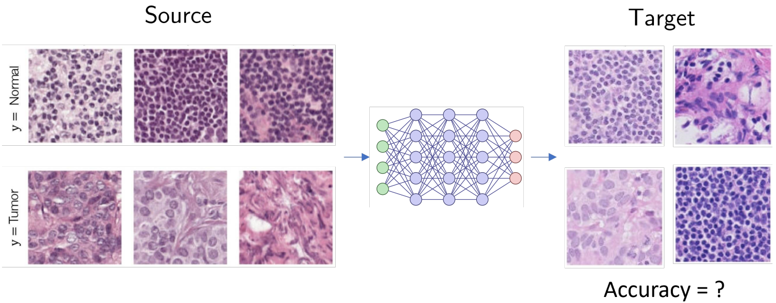




Dataset	Metric	In-dist setting	In-dist	Out-of-dist	Gap
iWildCam2020-wilds	Macro F1	Train-to-train	47.0 (1.4)	31.0 (1.3)	16.0
Camelyon17-wilds	Average acc	Train-to-train	93.2 (5.2)	70.3 (6.4)	22.9
RxRx1-wilds	Average acc	Mixed-to-test	39.8 (0.2)	29.9 (0.4)	9.9
OGB-MolPCBA	Average AP	Random split	34.4 (0.9)	27.2 (0.3)	7.2
GlobalWheat-wilds	Average domain acc	Mixed-to-test	63.3 (1.7)	49.6 (1.9)	13.7
CivilComments-wilds	Worst-group acc	Average	92.2 (0.1)	56.0 (3.6)	36.2
FMoW-wilds	Worst-region acc	Mixed-to-test	48.6 (0.9)	32.3 (1.3)	16.3
PovertyMap-wilds	Worst-U/R Pearson R	Mixed-to-test	0.60 (0.06)	0.45 (0.06)	0.15
Amazon-wilds	10th percentile acc	Average	71.9 (0.1)	53.8 (0.8)	18.1
Py150-wilds	Method/class acc	Train-to-train	75.4 (0.4)	67.9 (0.1)	7.5

Estimate OOD Performance

Can we estimate OOD performance?



Source

y = Normal

y = Tumor

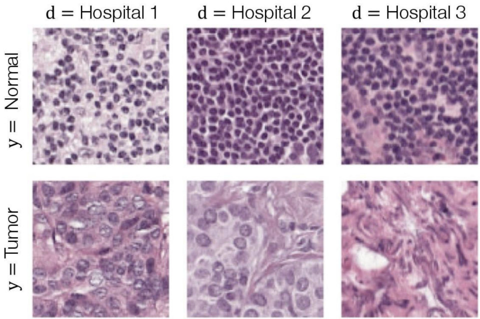
Target

Accuracy = ?

Why unlabeled data?

Often, we don’t have access to labeled data from target.

- It is expensive to label new data.



Why unlabeled data?

Often, we don’t have access to labeled data from target.

Why unlabeled data?

Often, we don't have access to labeled data from target.

- Rapidly changing environment. No time to label data.

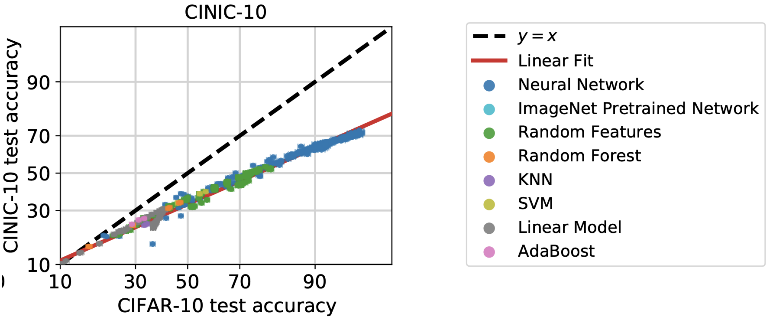
Challenge

Challenge:

Estimate out-of-distribution performance of a model only with unlabeled data.

Accuracy-on-the-line

- A linear trend between OOD and ID test error has been observed recently.



[Recht et al., 2019, Hendrycks et al., 2021, Koh et al., 2021, Taori et al., 2020, Miller et al., 2021]

Disagreement

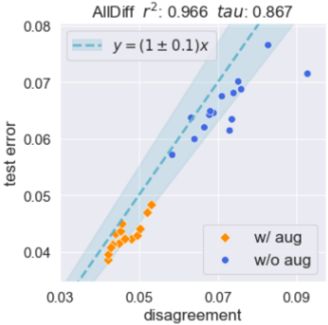
- Disagreement $= \mathbb{E}[\mathbf{1}\{f_1(x) \neq f_2(x)\}]$

Disagreement

- Disagreement $= \mathbb{E}[\mathbf{1}\{f_1(x) \neq f_2(x)\}]$
- Test disagreement $\approx$ test error

Disagreement

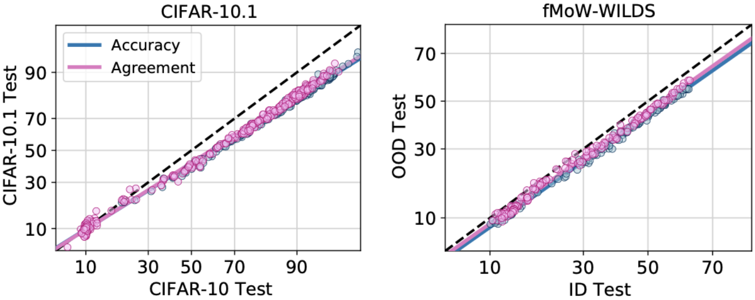
- Disagreement $= \mathbb{E}[\mathbf{1}\{f_1(x) \neq f_2(x)\}]$
- Test disagreement $\approx$ test error



[Hacohen et al., 2020, Chen et al., 2021, Jiang et al., 2021, Nakkiran and Bansal, 2020, Baek et al., 2022, Atanov et al., 2022, Pliushch et al., 2022]

Disagreement-on-the-line

- In particular, [Baek et al., 2022] found that OOD vs. ID agreement forms a line, and it closely matches that of the accuracy.



Introduction

We ask the following questions:

Introduction

We ask the following questions:

- Is disagreement-on-the-line a universal phenomenon?

Introduction

We ask the following questions:

- Is disagreement-on-the-line a universal phenomenon?
- Under what conditions is it guaranteed to happen?

Introduction

We ask the following questions:

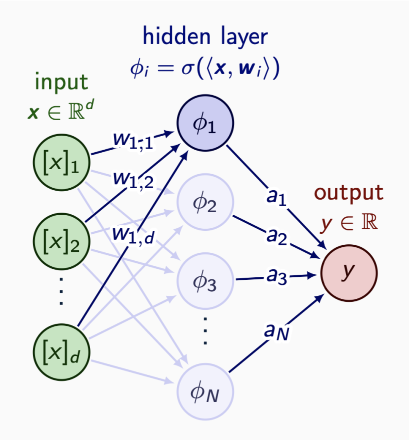
- Is disagreement-on-the-line a universal phenomenon?
- Under what conditions is it guaranteed to happen?
- What happens if those conditions fail?

Theoretical Model

Two-Layer Neural Network

$$f_{\text{NN}}(x) = a^\top \sigma\left(Wx\right)$$

- **Input data**: $x \in \mathbb{R}^d$.
- **Trainable parameters**: $W \in \mathbb{R}^{N \times d}, a \in \mathbb{R}^N$.
- **Nonlinearity**: $\sigma : \mathbb{R} \to \mathbb{R}$.



Problem Setting

Problem Setting

- **Data generation:**

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \quad y_i = f^\star(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2).$$

The input distribution *shifts* to $x \sim \mathsf{N}(0, \Sigma_{\mathrm{t}})$ at test time.

Problem Setting

- **Data generation:**

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \quad y_i = f^\star(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2).$$

  The input distribution *shifts* to $x \sim \mathsf{N}(0, \Sigma_{\mathrm{t}})$ at test time.

- **Ridge regression:**

$$\min_{W,a} \left[\frac{1}{n} \sum_{i=1}^{n} \Big(y_i - a^\top \sigma\left(W x_i\right) \Big)^2 + \gamma \|a\|_2^2 \right]$$

Problem Setting

- **Data generation:**

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \quad y_i = f^\star(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2).$$

  The input distribution *shifts* to $x \sim \mathsf{N}(0, \Sigma_{\mathrm{t}})$ at test time.

- **Ridge regression:**

$$\min_{W, a} \left[\frac{1}{n} \sum_{i=1}^{n} \Big(y_i - a^\top \sigma\left(W x_i\right) \Big)^2 + \gamma \|a\|_2^2 \right]$$

- **Optimization**:

Problem Setting

- **Data generation:**

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \quad y_i = f^\star(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2).$$

  The input distribution *shifts* to $x \sim \mathsf{N}(0, \Sigma_{\mathrm{t}})$ at test time.

- **Ridge regression:**

$$\min_{W,a} \left[\frac{1}{n} \sum_{i=1}^{n} \Big(y_i - a^\top \sigma\left(W x_i\right) \Big)^2 + \gamma \|a\|_2^2 \right]$$

- **Optimization**:
  - Optimizing a under fixed W is convex.

Problem Setting

- **Data generation:**

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \quad y_i = f^{\star}(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_{\varepsilon}^2).$$

  The input distribution *shifts* to $x \sim \mathsf{N}(0, \Sigma_{\mathrm{t}})$ at test time.

- **Ridge regression:**

$$\min_{W, a} \left[\frac{1}{n} \sum_{i=1}^{n} \left(y_i - a^{\top} \sigma \left(W x_i \right) \right)^2 + \gamma \|a\|_2^2 \right]$$

- **Optimization**:
  - Optimizing a under fixed W is convex.
  - Optimizing W under fixed a is non-convex.

Random Features Model

Random Features Model: We assume that the weight matrix W is randomly chosen and is not trained.

Random Features Model

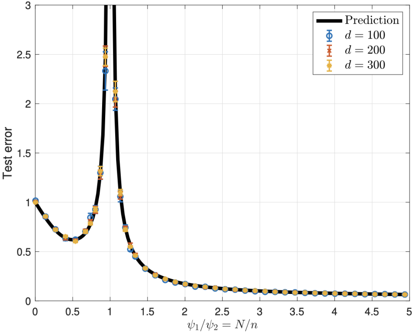
Random Features Model: We assume that the weight matrix W is randomly chosen and is not trained.

- If $n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$ (**proportional regime**), the risk will converge to a deterministic function of ϕ and ψ which can be computed.

Random Features Model

Random Features Model: We assume that the weight matrix W is randomly chosen and is not trained.

- If $n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$ (**proportional regime**), the risk will converge to a deterministic function of ϕ and ψ which can be computed.



Why random features model?

Why random features model?

It is able to capture various properties of neural networks:

Why random features model?

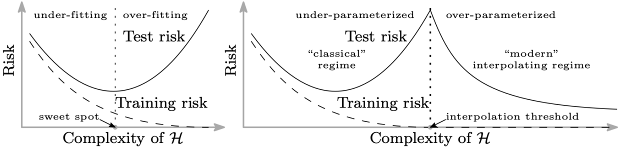
It is able to capture various properties of neural networks:

- Double Descent. [Belkin et al., 2019, Mei and Montanari, 2022]

Why random features model?

It is able to capture various properties of neural networks:

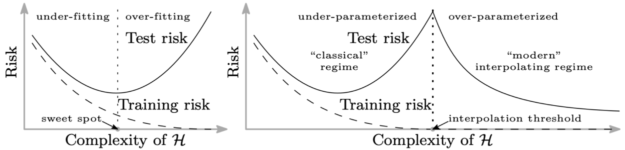
- Double Descent. [Belkin et al., 2019, Mei and Montanari, 2022]

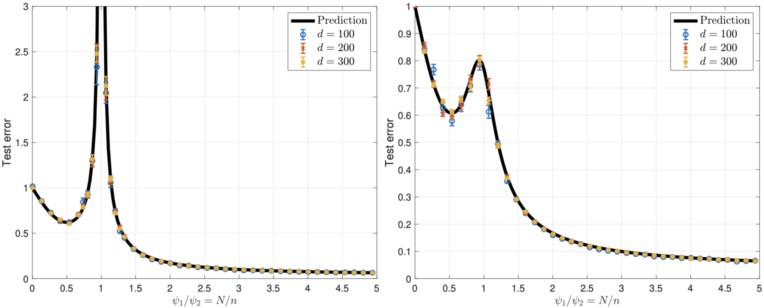


Why random features model?

It is able to capture various properties of neural networks:

- **Double Descent.** [Belkin et al., 2019, Mei and Montanari, 2022]





Why random features model?

It is able to capture various properties of neural networks:

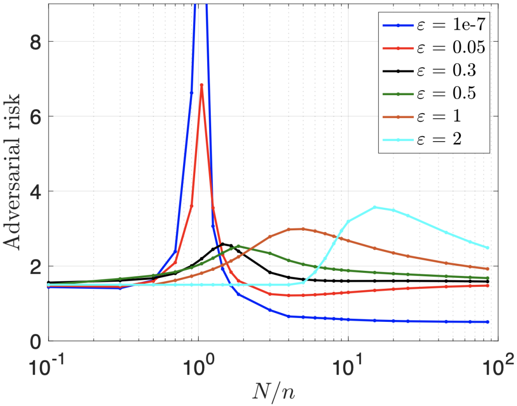
- Effect of overparameterization in adversarial training.
  [Hassani and Javanmard, 2022]

Why random features model?

It is able to capture various properties of neural networks:

- Effect of overparameterization in adversarial training.

  [Hassani and Javanmard, 2022]



Why random features model?

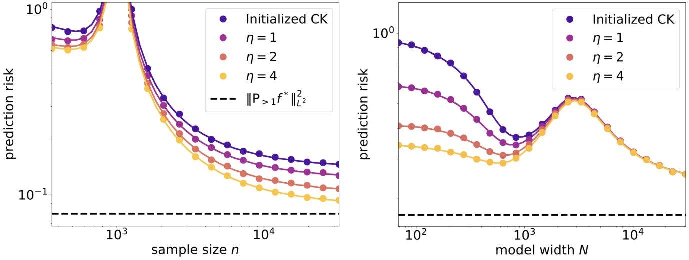
It is able to capture various properties of neural networks:

- Feature Learning. [Ba et al., 2022]

Why random features model?

It is able to capture various properties of neural networks:

- Feature Learning. [Ba et al., 2022]



(a) One step (risk vs. sample size).

(b) One step (risk vs. width).

Analysis of Disagreement

What is disagreement?

- We train two models with different “randomness" and then see how much they disagree on new data points.

What is disagreement?

- We train two models with different “randomness" and then see how much they disagree on new data points.
- Where does the randomness in training come from?

What is disagreement?

- We train two models with different “randomness" and then see how much they disagree on new data points.
- Where does the randomness in training come from?
  - Training set can be different.

What is disagreement?

- We train two models with different “randomness" and then see how much they disagree on new data points.
- Where does the randomness in training come from?
  - Training set can be different.
  - Weight initialization can be different.

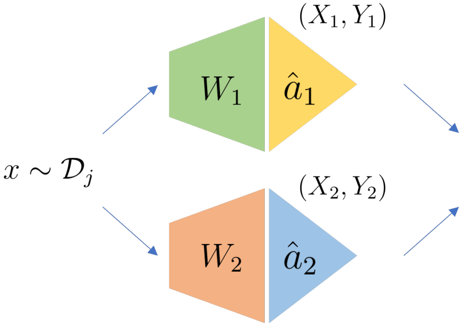
What is disagreement?

- We train two models with different “randomness" and then see how much they disagree on new data points.
- Where does the randomness in training come from?
  - Training set can be different.
  - Weight initialization can be different.
  - Order of mini-batches in training can be different.

What is disagreement?

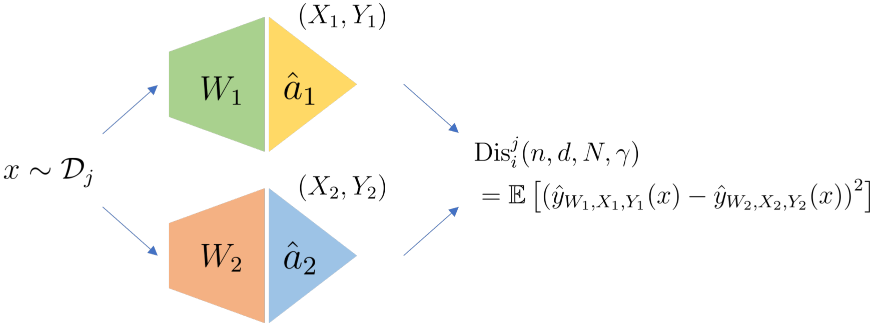
- We train two models with different “randomness" and then see how much they disagree on new data points.
- Where does the randomness in training come from?
  - Training set can be different.
  - Weight initialization can be different.
  - Order of mini-batches in training can be different.
- What do these translate to in random features models?

Types of Disagreement



$$\mathrm{Dis}_i^j(n, d, N, \gamma)$$
$$= \mathbb{E}\left[\left(\hat{y}_{W_1,X_1,Y_1}(x) - \hat{y}_{W_2,X_2,Y_2}(x)\right)^2\right]$$

Types of Disagreement



(X_1, Y_1)

W_1 $\hat{a}_1$

$x \sim \mathcal{D}_j$

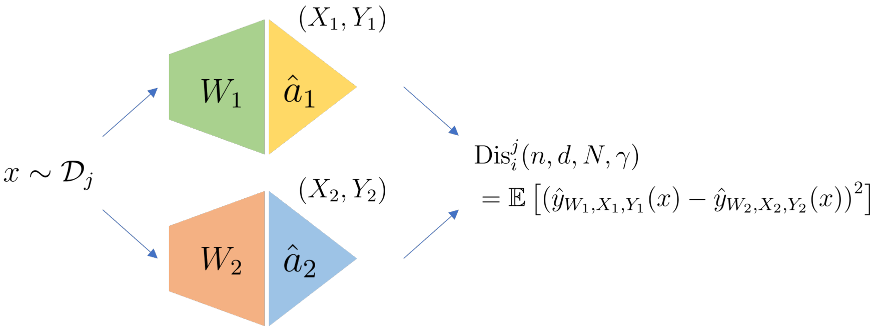
(X_2, Y_2)

W_2 $\hat{a}_2$

$$\mathrm{Dis}_i^j(n, d, N, \gamma) = \mathbb{E}\left[\left(\hat{y}_{W_1,X_1,Y_1}(x) - \hat{y}_{W_2,X_2,Y_2}(x)\right)^2\right]$$

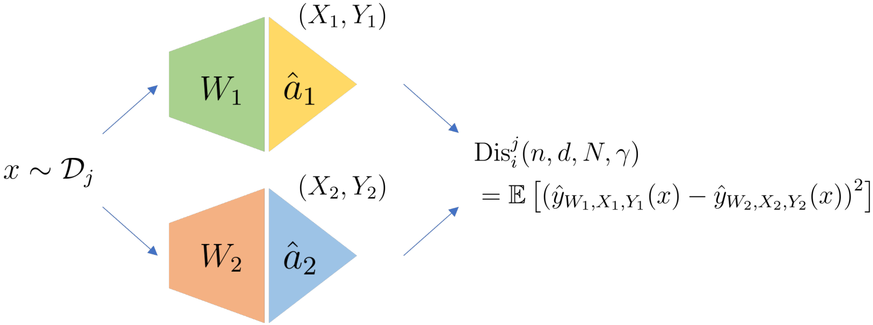
- **Independent** ($i = \mathrm{I}$): $(X_1, Y_1) \perp\!\!\!\perp (X_2, Y_2)$ and $W_1 \perp\!\!\!\perp W_2$.

Types of Disagreement


(X_1, Y_1)
W_1
$\hat{a}_1$
$x \sim \mathcal{D}_j$
(X_2, Y_2)
W_2
$\hat{a}_2$
$\mathrm{Dis}_i^j(n, d, N, \gamma)$
$= \mathbb{E}\left[\left(\hat{y}_{W_1,X_1,Y_1}(x) - \hat{y}_{W_2,X_2,Y_2}(x)\right)^2\right]$

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

Types of Disagreement



$$\mathrm{Dis}_i^j(n, d, N, \gamma) = \mathbb{E}\left[\left(\hat{y}_{W_1,X_1,Y_1}(x) - \hat{y}_{W_2,X_2,Y_2}(x)\right)^2\right]$$

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

Main Result

Theorem

Main Result

Theorem

In the proportional regime ($n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$), we derive exact asymptotic formulae for disgreements. For $j \in \{s, t\}$, we derive

$$\begin{aligned}\mathrm{Dis}_I^j(n, d, N, \gamma) &\to \delta_I^j(\phi, \psi, \gamma)\\ \mathrm{Dis}_{SS}^j(n, d, N, \gamma) &\to \delta_{SS}^j(\phi, \psi, \gamma)\\ \mathrm{Dis}_{SW}^j(n, d, N, \gamma) &\to \delta_{SW}^j(\phi, \psi, \gamma)\end{aligned}$$

Main Result

Theorem

In the proportional regime ($n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$), we derive exact asymptotic formulae for disgreements. For $j \in \{s, t\}$, we derive

$$\mathrm{Dis}_I^j(n, d, N, \gamma) \to \delta_I^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SS}^j(n, d, N, \gamma) \to \delta_{SS}^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SW}^j(n, d, N, \gamma) \to \delta_{SW}^j(\phi, \psi, \gamma)$$

The functions simplify a lot when we also take $\gamma \to 0$ (ridgeless).

Main Result

Theorem

In the proportional regime ($n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$), we derive exact asymptotic formulae for disgreements. For $j \in \{s, t\}$, we derive

$$\mathrm{Dis}_I^j(n, d, N, \gamma) \to \delta_I^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SS}^j(n, d, N, \gamma) \to \delta_{SS}^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SW}^j(n, d, N, \gamma) \to \delta_{SW}^j(\phi, \psi, \gamma)$$

The functions simplify a lot when we also take $\gamma \to 0$ (ridgeless).

[Please refer to Theorem 3.1 and Corollary 3.2 of the paper for the actual expressions.]

Main Result

Theorem

In the proportional regime ($n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$), we derive exact asymptotic formulae for disgreements. For $j \in \{s, t\}$, we derive

$$\mathrm{Dis}_I^j(n, d, N, \gamma) \to \delta_I^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SS}^j(n, d, N, \gamma) \to \delta_{SS}^j(\phi, \psi, \gamma)$$
$$\mathrm{Dis}_{SW}^j(n, d, N, \gamma) \to \delta_{SW}^j(\phi, \psi, \gamma)$$

The functions simplify a lot when we also take $\gamma \to 0$ (ridgeless).

[Please refer to Theorem 3.1 and Corollary 3.2 of the paper for the actual expressions.]

Proof technique: justify that the activation function can be linearized (Gaussian equivalence), then write disagreement as the expected trace of product of multiple random matrices and use tools from random matrix theory to analyze its limit.

Main Result

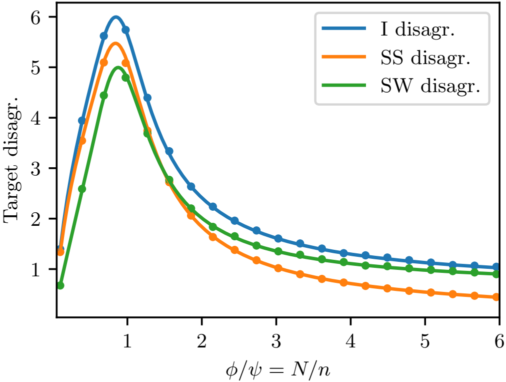


Figure: Asymptotics vs. simulations. We set $\gamma = 0.01$, $\sigma_\epsilon^2 = 0.25$, $\phi = 0.5$, $d = 512$, $n = 1024$, and specific Σ_s and Σ_t.

Linear Relation

Theorem (Exact linear relation)

Linear Relation

Theorem (Exact linear relation)

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

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

where the slopes and the intercepts are independent of ψ.

Linear Relation

Theorem (Exact linear relation)

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

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

where the slopes and the intercepts are independent of ψ.

The slopes and intercept depend on a specific measure of alignment between the source and target distribution.

Linear Relation

Theorem (Exact linear relation)

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

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

where the slopes and the intercepts are independent of ψ.

The slopes and intercept depend on a specific measure of alignment between the source and target distribution.

[Please refer to Theorem 4.1 of the paper for details.]

Linear Relation

Theorem (Exact linear relation)

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

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

where the slopes and the intercepts are independent of ψ.

The slopes and intercept depend on a specific measure of alignment between the source and target distribution.

[Please refer to Theorem 4.1 of the paper for details.]

The slope a is same as the slope for OOD vs. ID risk derived in [Tripuraneni et al., 2021] for the risk.

Linear Relation

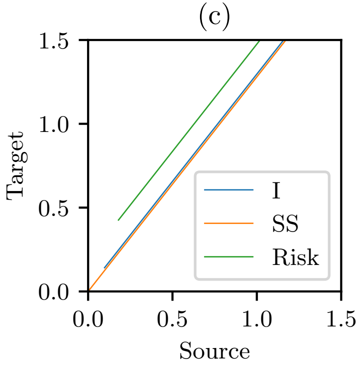


Figure: I and SS disagreement in the overparameterized and ridgeless setting have a linear relationship.

Linear Relation

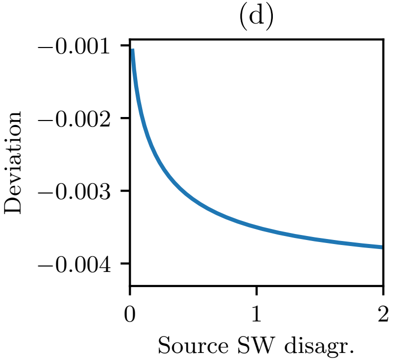


Figure: SW disagreement deviates from a line, even in the overparameterized and ridgeless setting.

Linear Relation

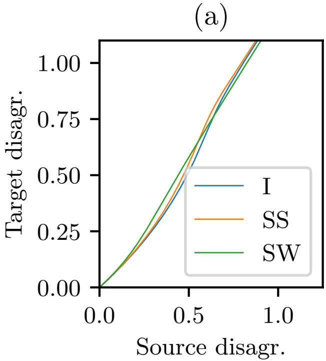


Figure: Disagreement-on-the-line *does not hold* in the underparametrized regime.

Approximate Linear Relation

Theorem (Approximate linear relation)

When $\gamma \neq 0$, in the overparameterized regime, deviation from the line, for I and SS disagreement, is bounded by

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

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

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

Approximate Linear Relation

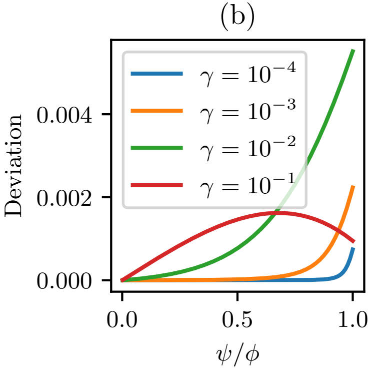


Figure: Deviation from the line, $\mathrm{Dis}^{\mathrm{t}}_{\mathrm{SS}}(\phi, \psi, \gamma) - a\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SS}}(\phi, \psi, \gamma)$, as a function of ψ for non-zero γ.

Summary

Table: Existence of disagreement-on-the-line in the overparameterized regime for different regularization and types of disagreement.

	Dis_{I} and Dis_{SS}		Dis_{SW}	
$\gamma \to 0$	✓	(Theorem 4.1)	✗	(Section 4.2)
$\gamma > 0$	▲	(Theorem 4.3)		

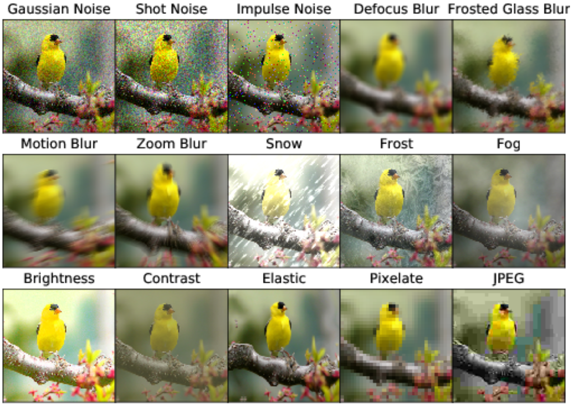
Experimental Results

Universality

- We proved everything for Gaussian input but read data is never Gaussian. Do our results still hold in more general settings?

Universality

- We proved everything for Gaussian input but read data is never Gaussian. Do our results still hold in more general settings?



Experiments

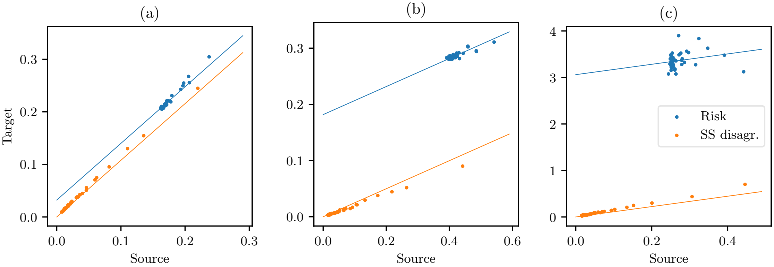


Figure: **(a)** CIFAR-10-C-Snow (severity 3) **(b)** Tiny ImageNet-C-Fog (severity 3) **(c)** Camelyon17; For more results, see Section D.3.

Experiments

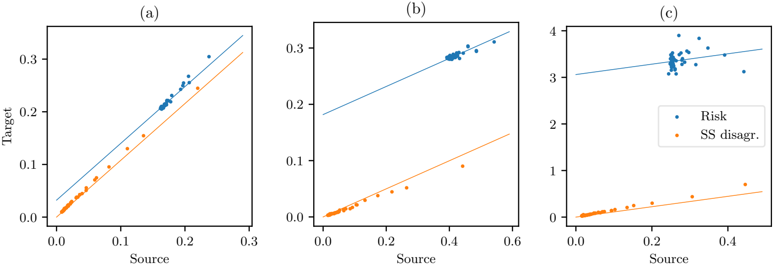


Figure: **(a)** CIFAR-10-C-Snow (severity 3) **(b)** Tiny ImageNet-C-Fog (severity 3) **(c)** Camelyon17; For more results, see Section D.3.

- Disagreement-on-the-line generalizes to non-Gaussian data (universality).

Experiments

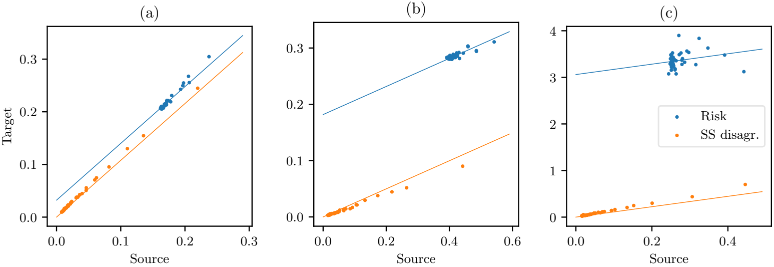


Figure: **(a)** CIFAR-10-C-Snow (severity 3) **(b)** Tiny ImageNet-C-Fog (severity 3) **(c)** Camelyon17; For more results, see Section D.3.

- Disagreement-on-the-line generalizes to non-Gaussian data (universality).
- Disagreement-on-the-line holds regardless of concept drift, i.e., the change in $\mathbb{P}(y|x)$.

Conclusion

Conclusion

- Disagreement-on-the-line is a nuanced phenomenon that can depend on the type of *randomness shared*, *regularization*, and the level of *overparametrization*.

Conclusion

- Disagreement-on-the-line is a nuanced phenomenon that can depend on the type of *randomness shared*, *regularization*, and the level of *overparametrization*.

- Contrary to the prior observation, the line for disagreement and the line for risk can *differ* in their intercepts.

Conclusion

- Disagreement-on-the-line is a nuanced phenomenon that can depend on the type of *randomness shared*, *regularization*, and the level of *overparametrization*.
- Contrary to the prior observation, the line for disagreement and the line for risk can *differ* in their intercepts.
- Experiments on several real-world datasets show that our theory is relevant beyond our theoretical setting.

Open Questions

- We showed that the intercepts for disagreement and risk can be different and characterized the difference. For practical use, can this difference in intercept be corrected?

Open Questions

- We showed that the intercepts for disagreement and risk can be different and characterized the difference. For practical use, can this difference in intercept be corrected?
- Extension to classification – challenges: no closed-form expression for $\hat{a}$, 0-1 risk cannot be expressed as a trace

Open Questions

- We showed that the intercepts for disagreement and risk can be different and characterized the difference. For practical use, can this difference in intercept be corrected?
- Extension to classification – challenges: no closed-form expression for $\hat{a}$, 0-1 risk cannot be expressed as a trace

Open Questions

- We showed that the intercepts for disagreement and risk can be different and characterized the difference. For practical use, can this difference in intercept be corrected?
- Extension to classification – challenges: no closed-form expression for $\hat{a}$, 0-1 risk cannot be expressed as a trace
- Training the weight matrix W – [Ba et al., 2022]

Open Questions

- We showed that the intercepts for disagreement and risk can be different and characterized the difference. For practical use, can this difference in intercept be corrected?
- Extension to classification – challenges: no closed-form expression for $\hat{a}$, 0-1 risk cannot be expressed as a trace
- Training the weight matrix W – [Ba et al., 2022]
- Formal justification of universality – [Montanari and Saeed, 2022, Goldt et al., 2022, Loureiro et al., 2021, Pesce et al., 2023]

References I

[Atanov et al., 2022] **Atanov, A., Filatov, A., Yeo, T., Sohmshetty, A., and Zamir, A. (2022).**
Task discovery: Finding the tasks that neural networks generalize on.
In *Advances in Neural Information Processing Systems*.

[Ba et al., 2022] **Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. (2022).**
High-dimensional asymptotics of feature learning: How one gradient step improves the representation.
In *Advances in Neural Information Processing Systems*.

[Baek et al., 2022] **Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. (2022).**
Agreement-on-the-line: Predicting the performance of neural networks under distribution shift.
In *Advances in Neural Information Processing Systems*.

[Belkin et al., 2019] **Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019).**
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences.

[Chen et al., 2021] **Chen, J., Liu, F., Avci, B., Wu, X., Liang, Y., and Jha, S. (2021).**
Detecting errors and estimating accuracy on unlabeled data with self-training ensembles.
In *Advances in Neural Information Processing Systems*.

[Goldt et al., 2022] **Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mézard, M., and Zdeborová, L. (2022).**
The Gaussian equivalence of generative models for learning with shallow neural networks.
In *Mathematical and Scientific Machine Learning*, pages 426–471.

[Hacohen et al., 2020] **Hacohen, G., Choshen, L., and Weinshall, D. (2020).**
Let's agree to agree: Neural networks share classification order on real datasets.
In *International Conference on Machine Learning*.

[Hassani and Javanmard, 2022] **Hassani, H. and Javanmard, A. (2022).**
The curse of overparametrization in adversarial training: Precise analysis of robust generalization for random features regression.
arXiv preprint arXiv:2201.05149.

References II

[Hendrycks et al., 2021] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. (2021).
The many faces of robustness: A critical analysis of out-of-distribution generalization.
In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

[Jiang et al., 2021] Jiang, Y., Nagarajan, V., Baek, C., and Kolter, J. Z. (2021).
Assessing generalization of SGD via disagreement.
In *International Conference on Learning Representations*.

[Koh et al., 2021] 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*.

[Loureiro et al., 2021] Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021).
Learning curves of generic features maps for realistic datasets with a teacher-student model.
In *Advances in Neural Information Processing Systems*.

[Mei and Montanari, 2022] Mei, S. and Montanari, A. (2022).
The generalization error of random features regression: Precise asymptotics and the double descent curve.
Communications on Pure and Applied Mathematics, 75(4):667–766.

[Miller et al., 2021] Miller, J. P., 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.
In *International Conference on Machine Learning*.

[Montanari and Saeed, 2022] Montanari, A. and Saeed, B. N. (2022).
Universality of empirical risk minimization.
In *Conference on Learning Theory*.

References III

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

[Pesce et al., 2023] Pesce, L., Krzakala, F., Loureiro, B., and Stephan, L. (2023).
Are gaussian data all you need? extents and limits of universality in high-dimensional generalized linear estimation.
arXiv preprint arXiv:2302.08923.

[Pliushch et al., 2022] Pliushch, I., Mundt, M., Lupp, N., and Ramesh, V. (2022).
When deep classifiers agree: Analyzing correlations between learning order and image statistics.
In *European Conference on Computer Vision*.

[Recht et al., 2019] Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019).
Do ImageNet classifiers generalize to ImageNet?
In *International Conference on Machine Learning*.

[Taori et al., 2020] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., and Schmidt, L. (2020).
Measuring robustness to natural distribution shifts in image classification.
In *Advances in Neural Information Processing Systems*.

[Tripuraneni et al., 2021] Tripuraneni, N., Adlam, B., and Pennington, J. (2021).
Overparameterization improves robustness to covariate shift in high dimensions.
In *Advances in Neural Information Processing Systems*.

Thank You!