Demystifying Disagreement-on-the-Line in High Dimensions

Donghwan Lee [*1] **Behrad Moniri** [*2] **Xinmeng Huang** [1] **Edgar Dobriban** [3] **Hamed Hassani** [2]

Abstract

Evaluating the performance of machine learning models under distribution shifts is challenging, especially when we only have unlabeled data from the shifted (target) domain, along with labeled data from the original (source) domain. Recent work suggests that the notion of *disagreement*, the degree to which two models trained with different randomness differ on the same input, is a key to tackling this problem. Experimentally, disagreement and prediction error have been shown to be strongly connected, which has been used to estimate model performance. Experiments have led to the discovery of the *disagreement-on-the-line* phenomenon, whereby the classification error under the target domain is often a linear function of the classification error under the source domain; and whenever this property holds, disagreement under the source and target domain follow the same linear relation. In this work, we develop a theoretical foundation for analyzing disagreement in high-dimensional random features regression; and study under what conditions the disagreement-on-the-line phenomenon occurs in our setting. Experiments on CIFAR-10-C, Tiny ImageNet-C, and Camelyon17 are consistent with our theory and support the universality of the theoretical findings.

1. Introduction

Modern machine learning methods such as deep neural networks are effective at prediction tasks when the input test data is similar to the data used during training. However, they can be extremely sensitive to changes in the input data distribution (e.g., Biggio et al. (2013); Szegedy et al. (2014); Hendrycks et al. (2020), etc.). This is a significant concern in safety-critical applications where errors are costly (e.g., Oakden-Rayner et al. (2020), etc.). In such scenarios, it is important to estimate how well the predictive model performs on out-of-distribution (OOD) data.

Collecting labeled data from new distributions can be costly, but unlabeled data is often readily available. As such, recent research efforts have focused on developing methods that can estimate a predictive model's OOD performance using only unlabeled data (e.g., Garg et al. (2021); Deng & Zheng (2021); Chen et al. (2021); Guillory et al. (2021), etc.).

In particular, works dating back at least to Recht et al. (2019) suggest that the out-of-distribution (OOD) and in-distribution (ID) errors of predictive models of different complexities are highly correlated. This was rigorously proved in Tripuraneni et al. (2021) for random features model under covariate shift. However, determining the correlation requires labeled OOD data. To sidestep this requirement, Baek et al. (2022) proposed an alternative approach that looks at the *disagreement* on an unlabeled set of data points between pairs of neural networks with the same architecture trained with different sources of randomness. They observed a linear trend between ID and OOD disagreement, as for ID and OOD error. Surprisingly, the linear trend *had the same empirical slope and intercept* as the linear trend between ID and OOD accuracy. This phenomenon, termed *disagreement-on-the-line*, allows estimating the linear relationship between OOD and ID error using only unlabeled data, and finally allows estimating the OOD error.

At the moment, the theoretical basis for disagreement-on-the-line remains unclear. It is unknown how generally it occurs, and what factors (such as the type of models or data used) may influence it. To better understand—or even demystify—these empirical findings, in this paper, we develop a theoretical foundation for studying disagreement. We focus on the following key questions:

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

To work towards answering these questions, we study dis-

[*]Equal contribution [1]Graduate Group in Applied Mathematics and Computational Science, University of Pennsylvania, PA, USA [2]Department of Electrical and Systems Engineering, University of Pennsylvania, PA, USA [3]Department of Statistics and Data Science, University of Pennsylvania, PA, USA. Correspondence to: Donghwan Lee <dh7401@sas.upenn.edu>, Behrad Moniri <bemoniri@seas.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

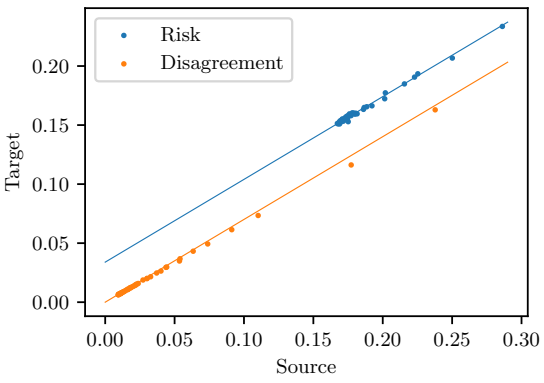

Figure 1. Target vs. source risk and shared-sample disagreement of random features model trained on CIFAR-10. Solid lines are derived from Theorem 4.1. Target domain is CIFAR-10-C-Fog (Hendrycks & Dietterich, 2018). See Section 5 for details.

agreement in a widely used theoretical framework for high-dimensional learning, *random features models*. We consider a setting where input data is from a Gaussian distribution, but possibly with a different covariance structure at training and test time, and study disagreement under the high-dimensional/proportional limit setting. We define various types of disagreement depending on what randomness the two models share. We *rigorously prove that depending on the type of shared randomness and the regime of parameterization, the disagreement-on-the-line may or may not happen in random feature models* trained using ridgeless least squares. Moreover, in contrast to prior observations, the line for disagreement and the line for risk may have *different intercepts*, even if they share the same slope. Additionally, we prove that adding ridge regularization breaks the exact linear relation, but an approximate linear relation still exists. Thus, we find that even in a simple theoretical setting, disagreement-on-the-line is a nuanced phenomenon that can depend on the type of randomness shared, regularization, and the level of overparametrization.

Experiments we performed on CIFAR-10-C and other datasets are consistent with our theory, even though the assumptions of Gaussianity of inputs and linearity of the data generation are not met (Figure 1, 4). This suggests that our theory is relevant beyond our theoretical setting.

1.1. Main Contributions

We provide an overview of the paper and our results.

- We propose a framework for the theoretical study of disagreement. We introduce a comprehensive and unifying set of notions of disagreement (Definition 2.1). Then, we find a limiting formula for disagreement in the high-dimensional limit where the sample size, input dimension, and feature dimension grow proportionally (Theorem 3.1).
- Based on this characterization, we study how disagreement under source and target domains are related. We identify under what conditions and for which type of disagreement the *disagreement-on-the-line* phenomenon holds (Section 4). Theorem 4.3 and Corollary 4.4 show an approximate linear relation when the conditions are not met.
- When the disagreement-on-the-line holds in our model, our results imply that the *target vs. source line for risk* and the *target vs. source line for disagreement* have the same slope. This is consistent with the findings of Baek et al. (2022), that whenever OOD vs. ID accuracy is on a line, OOD vs. ID agreement is also on the same line. However, unlike their finding, in our problem, the intercepts of the lines can be different (Remark 4.2).
- In Section 5, we conduct experiments on several datasets including CIFAR-10-C, Tiny ImageNet-C, and Camelyon17. The experimental results are generally consistent with our theoretical findings, even as the theoretical conditions we use (e.g., Gaussian input, linear generative model, etc.) may not hold. This suggests a possible universality of the theoretical predictions.
- Our work shows that disagreement-on-the-line is a subtle phenomenon that depends on the shared randomness, regularization, and regime of parameterization. We also identify a difference between the intercept of the line for risk and the line for disagreement. If these factors are not properly considered, the disagreement-on-the-line principle can lead to an inaccurate OOD performance estimation.

1.2. Related Work

Random Features Model. Random features models were introduced by Rahimi & Recht (2007) as an approach for scaling kernel methods to massive datasets. Recently, they have been used as a standard model for the theoretical study of deep neural networks. Despite its simplicity, it is rich enough to capture various phenomena of deep learning including double descent (Mei & Montanari, 2022; Adlam et al., 2022; Lin & Dobriban, 2021), adversarial training (Hassani & Javanmard, 2022), feature learning (Ba et al., 2022), and transfer learning (Tripuraneni et al., 2021). In particular, in this model, the number of parameters and the ambient dimension are disentangled, hence the effect of overparameterization can be studied on its own.

Linear Relation Under Distribution Shift. Several intriguing phenomena have been observed in empirical studies of distribution shifts. Recht et al. (2019); Hendrycks et al.

(2021); Koh et al. (2021); Taori et al. (2020); Miller et al. (2021) observed linear trends between OOD and ID test error. Tripuraneni et al. (2021) proved this phenomenon in random features models under covariate shift.

Recently, the notion of disagreement has been gaining a lot of attention (e.g., Hacohen et al. (2020); Chen et al. (2021); Jiang et al. (2021); Nakkiran & Bansal (2020); Baek et al. (2022); Atanov et al. (2022); Pliushch et al. (2022), etc.).

In particular, Baek et al. (2022) empirically showed that OOD agreement between the predictions of pairs of neural networks also has a strong linear correlation with their ID agreement. They further observed that the slope and intercept of the OOD vs ID agreement line closely match that of the accuracy. This can be used to predict the OOD performance of predictive models only using unlabeled data.

High-dimensional Asymptotics. Work on high-dimensional asymptotics dates back at least to the 1960s (Raudys, 1967; Deev, 1970; Raudys, 1972) and has more recently been studied in a wide range of areas, such as high-dimensional statistics (e.g., Raudys & Young (2004); Serdobolskii (2007); Paul & Aue (2014); Yao et al. (2015); Dobriban & Wager (2018), etc.), wireless communications (e.g., Tulino & Verdú (2004); Couillet & Debbah (2011), etc.), and machine learning (e.g., Györgyi & Tishby (1990); Opper (1995); Opper & Kinzel (1996); Couillet & Liao (2022); Engel & Van den Broeck (2001), etc.).

Technical Tools. The results derived in this paper rely on the Gaussian equivalence conjecture studied and used extensively for random features model (e.g., Goldt et al. (2022); Hu & Lu (2022); Montanari & Saeed (2022); Mei & Montanari (2022); Hassani & Javanmard (2022); Tripuraneni et al. (2021); Loureiro et al. (2021); d'Ascoli et al. (2021), etc.). Our analytical results build upon the series of recent work Mel & Pennington (2021); Adlam & Pennington (2020a); Tripuraneni et al. (2021) using random matrix theory and operator-valued free probability (Far et al., 2008; Mingo & Speicher, 2017).

2. Preliminaries

2.1. Problem Setting

We study a supervised learning setting where the training data $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i \in [n]$, of dimension d and sample size n, is generated according to

$$x_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \Sigma_{\mathrm{s}}), \text{ and } y_i = \frac{1}{\sqrt{d}} \beta^\top x_i + \varepsilon_i, \qquad (1)$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, \sigma_\varepsilon^2)$. Additionally, the true coefficient $\beta \in \mathbb{R}^d$ is assumed to be randomly drawn from $\mathsf{N}(0, I_d)$. The linear relationship between (x_i, y_i) is not known. We fit a model to the data, which can then be used to predict labels for unlabeled examples at test time.

We consider two-layer neural networks with fixed, randomly generated weights in the first layer—a random features model—as the learner. We let the width of the internal layer be $N \in \mathbb{N}$. For a weight matrix $W \in \mathbb{R}^{N \times d}$ with i.i.d. random entries sampled from $\mathsf{N}(0, 1)$, an activation function $\sigma : \mathbb{R} \to \mathbb{R}$ applied elementwise, and the weights $a \in \mathbb{R}^N$ of a linear layer, the random features model is defined by

$$f_{W,a}(x) = \frac{1}{\sqrt{N}} a^\top \sigma\left(Wx/\sqrt{d}\right).$$

The trainable parameters $a \in \mathbb{R}^N$ are fit via ridge regression to the training data $X = (x_1, \ldots, x_n) \in \mathbb{R}^{d \times n}$ and $Y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Specifically, for a regularization parameter $\gamma > 0$, we solve

$$\hat{a} = \underset{a \in \mathbb{R}^N}{\arg\min} \left\| Y - \sigma\left(WX/\sqrt{d}\right)^\top a/\sqrt{N} \right\|_2^2 + \gamma \|a\|_2^2,$$

and use $\hat{y}(x) = \hat{a}^\top \sigma(Wx/\sqrt{d})/\sqrt{N}$ as the model prediction for a data point $x \in \mathbb{R}^d$. Defining $F = \sigma(WX/\sqrt{d})$ and $f = \sigma(Wx/\sqrt{d})$, we can write

$$\hat{y}(x) = Y^\top \left(\frac{1}{N} F^\top F + \gamma I_n\right)^{-1} \left(\frac{1}{N} F^\top f\right). \qquad (2)$$

To emphasize the dependence on W, X, Y, we also use the notation $\hat{y}_{W,X,Y}$.

It has been recognized in e.g., Adlam & Pennington (2020a); Ghorbani et al. (2021); Mei & Montanari (2022) that only linear data generative models can be learned in the proportional-limit high-dimensional regime by random features models, and the non-linear part behaves like an additive noise. Thus, we consider linear generative models as in (1). Results for non-linear models can be obtained via linearization, as is standard in the above work.

We also highlight that our theoretical findings are validated by simulations on standard datasets (such as CIFAR-10-C) where the input distribution is non-Gaussian and the data generation model is non-linear.

2.2. Distribution Shift

At training time (1), the inputs x_i are sampled from the *source domain*, $\mathcal{D}_{\mathrm{s}} = \mathsf{N}(0, \Sigma_{\mathrm{s}})$. At test time, we assume the input distribution shifts to the *target domain*, $\mathcal{D}_{\mathrm{t}} = \mathsf{N}(0, \Sigma_{\mathrm{t}})$. We do not restrict the change in $\mathbb{P}(y|x)$ since disagreement is independent of the label y. Previous work (Lei et al., 2021; Tripuraneni et al., 2021; Wu et al., 2022) found that the learning problem under covariate shift is fully characterized by input covariance matrices. For this reason, we do not consider shifts in the mean of the input distribution.

2.3. Definition of Disagreement

Hacohen et al. (2020); Chen et al. (2021); Jiang et al. (2021); Nakkiran & Bansal (2020); Baek et al. (2022) define notions of *disagreement* (or *agreement*) to quantify the difference (or similarity) between the predictions of two randomly trained predictive models in *classification* tasks.

Prior work on disagreement considers three sources of randomness that lead to different predictive models: (i) random initialization, (ii) sampling of the training set, and (iii) sampling/ordering of mini-batches.

Motivated by these results, we propose analogous notions of disagreement in random features regression. We consider (i), (ii) and their combination, as (iii) is not present in our problem.

The *independent disagreement* measures how much the prediction of two models with independent random weights and trained on two independent sets of training datasets disagree, on average. Similar notions were used in (Nakkiran & Bansal, 2020; Pliushch et al., 2022; Jiang et al., 2021; Baek et al., 2022).

The *shared-sample disagreement* measures the average difference of the predictions of two models with independent random weights, but trained on a shared training set. Similar notions were used in (Pliushch et al., 2022; Jiang et al., 2021; Baek et al., 2022; Atanov et al., 2022).

The *shared-weight disagreement* measures the average difference of the predictions of two models with shared random weights, but trained on two independent training samples. Similar notions were used in (Jiang et al., 2021; Baek et al., 2022).

While the prior work typically used 0-1 loss to define agreement/disagreement in classification, we use the squared loss to measure disagreement for real-valued outputs.

Definition 2.1 (Disagreement)**.** Consider two random features models trained on the data $(X_1, Y_1), (X_2, Y_2) \in \mathbb{R}^{d\times n} \times \mathbb{R}^n$ with random weight matrices $W_1, W_2 \in \mathbb{R}^{N\times d}$, respectively. We measure the disagreement of two models by their mean squared difference

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

where the expectation is over $\beta, W_1, W_2, X_1, Y_1, X_2, Y_2$, and $j \in \{\mathrm{s}, \mathrm{t}\}$ is the domain that $x \sim \mathcal{D}_j$ is from, and the index $i \in \{\mathrm{I}, \mathrm{SS}, \mathrm{SW}\}$ corresponds to one of the following cases.

- **I**ndependent disagreement ($i = \mathrm{I}$): the training data $(X_1, Y_1), (X_2, Y_2)$ are independently generated from (1), with the same β. The weights $W_1, W_2 \in \mathbb{R}^{N\times d}$ are independent matrices with i.i.d. $\mathsf{N}(0, 1)$ entries.
- **S**hared-**S**ample disagreement ($i = \mathrm{SS}$): the training samples are shared, i.e., $(X_1, Y_1) = (X_2, Y_2) = (X, Y)$, where (X, Y) is generated from (1). The weights $W_1, W_2 \in \mathbb{R}^{N\times d}$ are independent matrices with i.i.d. $\mathsf{N}(0, 1)$ entries.
- **S**hared-**W**eight disagreement ($i = \mathrm{SW}$): the training data $(X_1, Y_1), (X_2, Y_2)$ are independently generated from (1), with the same β. Two models share the weights, i.e., $W_1 = W_2 = W$. The weights are shared, i.e., $W_1 = W_2 = W$, where $W \in \mathbb{R}^{N\times d}$ is a matrix with i.i.d. $\mathsf{N}(0, 1)$ entries.

2.4. Conditions

We characterize the asymptotics of disagreement in the proportional limit asymptotic regime defined as follows.

Condition 2.2 (Asymptotic setting)**.** We assume that $n, d, N \to \infty$ with $d/n \to \phi > 0$ and $d/N \to \psi > 0$.

To characterize the limit of disagreement, we need conditions on the spectral properties of Σ_s and Σ_t as their dimension d grows. When multiple growing matrices are involved, it is not sufficient to make assumptions on the individual spectra of the matrices, but rather, they have to be considered *jointly* (Wu & Xu, 2020; Tripuraneni et al., 2021; Mel & Pennington, 2021). We assume that the *joint spectral distribution* of Σ_s and Σ_t converges to a limiting distribution μ on $\mathbb{R}_+^2$ as $d \to \infty$.

Condition 2.3. Let $\lambda_1^\mathrm{s}, \ldots, \lambda_d^\mathrm{s} \geq 0$ be the eigenvalues of Σ_s and $v_1, \ldots, v_d$ be the corresponding eigenvectors. Define $\lambda_i^\mathrm{t} = v_i^\top \Sigma_\mathrm{t} v_i$ for $i \in [d]$. We assume the joint empirical spectral distribution of $(\lambda_i^\mathrm{s}, \lambda_i^\mathrm{t}), i \in [d]$ converges in distribution to a limiting distribution μ on $\mathbb{R}_+^2$. That is, $\frac{1}{d}\sum_{i=1}^d \delta_{(\lambda_i^\mathrm{s}, \lambda_i^\mathrm{t})} \to \mu$, where δ is the Dirac delta measure. We additionally assume that μ has a compact support. We denote random variables drawn from μ by $(\lambda^\mathrm{s}, \lambda^\mathrm{t})$, and write $m_\mathrm{s} = \mathbb{E}_\mu[\lambda^\mathrm{s}]$ and $m_\mathrm{t} = \mathbb{E}_\mu[\lambda^\mathrm{t}]$.

For the existence of certain derivatives and expectations, we assume the following mild condition on the activation function $\sigma : \mathbb{R} \to \mathbb{R}$.

Condition 2.4. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is differentiable almost everywhere. There are constants c_0 and c_1 such that $|\sigma(x)|, |\sigma'(x)| \leq c_0 e^{c_1 x}$, whenever $\sigma'(x)$ exists. For $j \in \{\mathrm{s}, \mathrm{t}\}$ and a standard Gaussian random variable $Z \sim \mathsf{N}(0, 1)$, define

$$\rho_j = \frac{\mathbb{E}[Z\sigma(\sqrt{m_j}Z)]^2}{m_j}, \; \omega_j = \frac{\mathbb{V}[\sigma(\sqrt{m_j}Z)]}{\rho_j} - m_j. \quad (3)$$

These constants characterize the non-linearity of the activation σ and will appear in the asymptotics of disagreement. Note that when σ is ReLU activation $\sigma(x) = \max(x, 0)$, we have $\rho_j = 1/4, \omega_j = m_j(1 - 2/\pi)$ for $j \in \{\mathrm{s}, \mathrm{t}\}$.

3. Asymptotics of Disagreement

In this section, we present our results on characterizing the limits of disagreement defined in Definition 2.1 for random features models. We introduce results for general ridge regression and also study the *ridgeless limit* $\gamma \to 0$.

3.1. Ridge Setting

For $i \in \{\mathrm{I}, \mathrm{SS}, \mathrm{SW}\}$ and $j \in \{\mathrm{s}, \mathrm{t}\}$, define the *asymptotic disagreement*

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

where the limit is in the regime considered in Condition 2.2.

Asymptotics in random features models and linear models with general covariance (e.g., training/test error, bias, variance, etc.) typically do not have a closed form, and can only be implicitly described through *self-consistent equations* (Tulino & Verdú, 2004; Dobriban & Wager, 2018; Adlam et al., 2022; Mei & Montanari, 2022; Hastie et al., 2022). To facilitate analysis of these implicit quantities, previous work (e.g., Dobriban & Sheng (2021; 2020); Tripuraneni et al. (2021); Mel & Pennington (2021), etc.) proposed using expressions containing *only one implicit scalar*. We show that similar to the asymptotic risk derived in Tripuraneni et al. (2021), the asymptotic disagreements can be expressed using a scalar κ which is the unique non-negative solution of the *self-consistent equation*

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

where $\mathcal{I}^j_{a,b}$ is the *integral functional* of μ defined by

$$\mathcal{I}^j_{a,b}(\kappa) = \phi\mathbb{E}_\mu\left[\frac{(\lambda^{\mathrm{s}})^{a-1}\lambda^j}{(\phi+\kappa\lambda^{\mathrm{s}})^b}\right], \quad j \in \{\mathrm{s}, \mathrm{t}\}. \tag{5}$$

We omit κ and simply write $\mathcal{I}^j_{a,b}$ whenever the argument is clear from the context. Recall from Condition 2.3 that μ describes the joint spectral properties of source and target covariance matrices, so $\mathcal{I}^j_{a,b}$ can be viewed as a summary of the joint spectral properties.

The following theorem—our first main result—shows that $\mathrm{Dis}^j_{\mathrm{I}}(\phi,\psi,\gamma)$, $\mathrm{Dis}^j_{\mathrm{SS}}(\phi,\psi,\gamma)$, $\mathrm{Dis}^j_{\mathrm{SW}}(\phi,\psi,\gamma)$ are well defined, and characterizes them.

Theorem 3.1 (Disagreement in general ridge regression)**.** *For $j \in \{\mathrm{s}, \mathrm{t}\}$, the asymptotic independent disagreement is*

$$\begin{aligned}
&\mathrm{Dis}^j_{\mathrm{I}}(\phi,\psi,\gamma) \\
&= \frac{2\rho_j\psi\kappa}{\phi\gamma + \rho_{\mathrm{s}}\gamma(\tau\psi+\bar\tau\phi)(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})}\Big[\gamma\tau(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\mathrm{s}}_{2,2} \\
&\quad + (\sigma_\varepsilon^2 + \mathcal{I}^{\mathrm{s}}_{1,1})(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})(\omega_j+\mathcal{I}^j_{1,1}) \\
&\quad + \frac{\phi}{\psi}\gamma\bar\tau(\sigma_\varepsilon^2+\phi\mathcal{I}^{\mathrm{s}}_{1,2})\mathcal{I}^j_{2,2}\Big],
\end{aligned}$$

and the asymptotic shared-sample disagreement is

$$\begin{aligned}
&\mathrm{Dis}^j_{\mathrm{SS}}(\phi,\psi,\gamma) \\
&\quad = \mathrm{Dis}^j_{\mathrm{I}}(\phi,\psi,\gamma) - \frac{2\rho_j\kappa^2(\sigma_\varepsilon^2+\phi\mathcal{I}^{\mathrm{s}}_{1,2})\mathcal{I}^j_{2,2}}{\rho_{\mathrm{s}}(1-\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})},
\end{aligned}$$

and the asymptotic shared-weight disagreement is

$$\begin{aligned}
&\mathrm{Dis}^j_{\mathrm{SW}}(\phi,\psi,\gamma) \\
&\quad = \mathrm{Dis}^j_{\mathrm{I}}(\phi,\psi,\gamma) - \frac{2\rho_j\psi\kappa^2(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\mathrm{s}}_{2,2}}{\rho_{\mathrm{s}}(\phi-\psi\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})},
\end{aligned}$$

where τ and $\bar\tau$ are the limiting normalized trace of $(F^\top F/N+\gamma I_n)^{-1}$ and $(FF^\top/N+\gamma I_N)^{-1}$, respectively. They can be expressed as functions of κ as follows:

$$\begin{aligned}
\tau &= \frac{\sqrt{(\psi-\phi)^2+4\kappa\psi\phi\gamma/\rho_{\mathrm{s}}}+\psi-\phi}{2\psi\gamma}, \\
\bar\tau &= \frac{1}{\gamma} + \frac{\psi}{\phi}\left(\tau - \frac{1}{\gamma}\right).
\end{aligned} \tag{6}$$

The expressions in Theorem 3.1 are written in terms of the non-linearity constants $\rho_{\mathrm{s}}, \rho_{\mathrm{t}}, \omega_{\mathrm{s}}, \omega_{\mathrm{t}}$, the dimension parameters ψ, ϕ, the regularization γ, the noise level σ_ε^2, the summary statistics $\mathcal{I}^{\mathrm{s}}_{a,b}, \mathcal{I}^{\mathrm{t}}_{a,b}$ of μ, and $\tau, \bar\tau, \kappa$. Since $\tau, \bar\tau$ are algebraic functions of κ, the expressions are functions of *one implicit variable* κ.

This theorem can be used to make numerical predictions for disagreement. To do so, we first solve the self-consistent equation (4) using a fixed-point iteration and find κ. Then, we plug κ into the terms appearing in the theorem. Figure 2 shows an example, supporting that the theoretical predictions of Theorem 3.1 match very well with simulations even for moderately large d, n, N.

Theoretical Innovations. To prove this theorem, we first rely on Gaussian equivalence (Section A.3, A.4) to express disagreement as a combination of traces of rational functions of i.i.d. Gaussian matrices. Then, we construct linear pencils (Section A.5) and use the theory of operator-valued free probability (Section A.1, A.2) to derive the limit of these trace objects. This general strategy has been used previously in Adlam et al. (2022); Adlam & Pennington (2020b); Tripuraneni et al. (2021); Mel & Pennington (2021).

However, in the expressions of disagreement, new traces appear that did not exist in prior work. We construct new suitable linear pencils to derive the limit of these traces. While this leads to a coupled system of self-consistent equations of many variables, it turns out that they can be factored into a single scalar variable κ defined through the self-consistent equation (4), and every term appearing in the limiting disagreements, can be written as algebraic functions

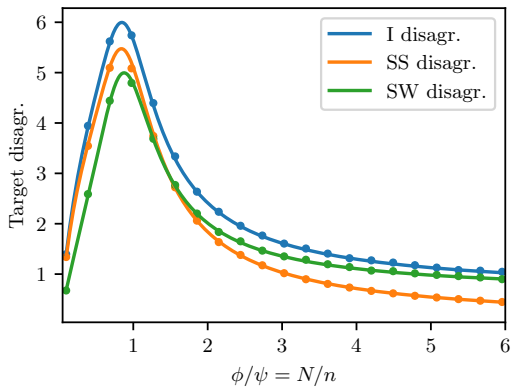

Figure 2. Independent, shared-sample, and shared-weight disagreement under target domain in random features regression with ReLU activation function, $\phi = \lim d/n = 0.5$, versus $\phi/\psi = \lim N/n$. We set $\gamma = 0.01$, $\sigma_\varepsilon^2 = 0.25$, and $\mu = 0.5\delta_{(1.5,5)} + 0.5\delta_{(1,1)}$. Simulations are done with $d = 512$, $n = 1024$, and averaged over 300 trials. The continuous lines are theoretical predictions from Theorem 3.1, and the dots are simulation results.

of κ. These results might also be of independent interest. Since limiting disagreements only rely on the same implicit variable as the variable appearing in the limiting risk, we can derive the results in Section 4.

3.2. Ridgeless Limit

In the ridgeless limit $\gamma \to 0$, the self-consistent equation (4) for κ becomes

$$\kappa = \frac{\min(1, \phi/\psi)}{\omega_{\mathrm{s}} + \mathcal{I}_{1,1}^{\mathrm{s}}(\kappa)}. \tag{7}$$

Further, the asymptotic limits in Theorem 3.1 can be simplified as follows.

Corollary 3.2 (Ridgeless limit)**.** *For $j \in \{\mathrm{s}, \mathrm{t}\}$ and in the ridgeless limit $\gamma \to 0$, the asymptotic independent disagreement is*

$$\lim_{\gamma\to 0} \mathrm{Dis}_{\mathrm{I}}^{j}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\mathrm{s}}|\phi-\psi|}(\sigma_\varepsilon^2 + \mathcal{I}_{1,1}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^{j}) + \begin{cases} \frac{2\rho_j\kappa(\sigma_\varepsilon^2+\phi\mathcal{I}_{1,2}^{\mathrm{s}})\mathcal{I}_{2,2}^{j}}{\rho_{\mathrm{s}}(\omega_{\mathrm{s}}+\phi\mathcal{I}_{1,2}^{\mathrm{s}})} & \phi > \psi, \\ \frac{2\rho_j\kappa(\omega_j+\phi\mathcal{I}_{1,2}^{j})\mathcal{I}_{2,2}^{\mathrm{s}}}{\rho_{\mathrm{s}}(\omega_{\mathrm{s}}+\phi\mathcal{I}_{1,2}^{\mathrm{s}})} & \phi < \psi, \end{cases}$$

and the asymptotic shared-sample disagreement is

$$\lim_{\gamma\to 0} \mathrm{Dis}_{\mathrm{SS}}^{j}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\mathrm{s}}|\phi-\psi|}(\sigma_\varepsilon^2 + \mathcal{I}_{1,1}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^{j}) + \begin{cases} 0 & \phi > \psi, \\ \frac{2\rho_j\kappa}{\rho_{\mathrm{s}}}\left(\frac{(\omega_j+\phi\mathcal{I}_{1,2}^{j})\mathcal{I}_{2,2}^{\mathrm{s}}}{\omega_{\mathrm{s}}+\phi\mathcal{I}_{1,2}^{\mathrm{s}}} - \frac{\kappa(\sigma_\varepsilon^2+\phi\mathcal{I}_{1,2}^{\mathrm{s}})\mathcal{I}_{2,2}^{j}}{1-\kappa^2\mathcal{I}_{2,2}^{\mathrm{s}}}\right) & \phi < \psi, \end{cases}$$

Table 1. Existence of disagreement-on-the-line in the overparametrized regime for different regularization and types of disagreement. The symbols ✓, ▲, ✗ correspond to exact, approximate, no linear relation, respectively.

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

and the asymptotic shared-weight disagreement is

$$\lim_{\gamma\to 0} \mathrm{Dis}_{\mathrm{SW}}^{j}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\mathrm{s}}|\phi-\psi|}(\sigma_\varepsilon^2 + \mathcal{I}_{1,1}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^{j}) + \begin{cases} \frac{2\rho_j\kappa}{\rho_{\mathrm{s}}}\left(\frac{(\sigma_\varepsilon^2+\phi\mathcal{I}_{1,2}^{\mathrm{s}})\mathcal{I}_{2,2}^{j}}{\omega_{\mathrm{s}}+\phi\mathcal{I}_{1,2}^{\mathrm{s}}} - \frac{\psi\kappa(\omega_j+\phi\mathcal{I}_{1,2}^{j})\mathcal{I}_{2,2}^{\mathrm{s}}}{\phi-\psi\kappa^2\mathcal{I}_{2,2}^{\mathrm{s}}}\right) & \phi > \psi, \\ 0 & \phi < \psi, \end{cases}$$

where κ is defined in (7).

In the ridgeless limit, I and SS disagreement have a *single term* that depends on ψ, which motivates the analysis in Section 4 that examines the *disagreement-on-the-line* phenomenon. In contrast, SW disagreement has two linearly independent terms that are functions of ψ, leading to a distinct behavior compared to I and SS disagreement.

The asymptotics in Corollary 3.2 reveal another interesting phenomenon regarding disagreements of random features model in the ridgeless limit. For example, it follows from Corollary 3.2 that SS disagreement tends to zero in the *infinite overparameterization* limit where the width N of the internal layer is much larger than the data dimension d, so that $\psi = \lim d/N \to 0$. However, the same is not true for the I and SW disagreement. This indicates that, in the infinite overparameterization limit, the randomness caused by the random weights disappears, and the model is solely determined by the training sample.

4. When Does Disagreement-on-the-Line Hold?

In this section, based on the characterizations of disagreements derived in the previous section, we study for which types of disagreement and under what conditions, the *linear relationship* between disagreement under source and target domain of models of varying complexity holds.

4.1. I and SS disagreement

Ridgeless. In the overparametrized regime $\phi > \psi$, the self-consistent equation (7) is independent of $\psi = \lim d/N$, and so is κ. This implies the following linear trend of I and SS disagreement, in the ridgeless limit.

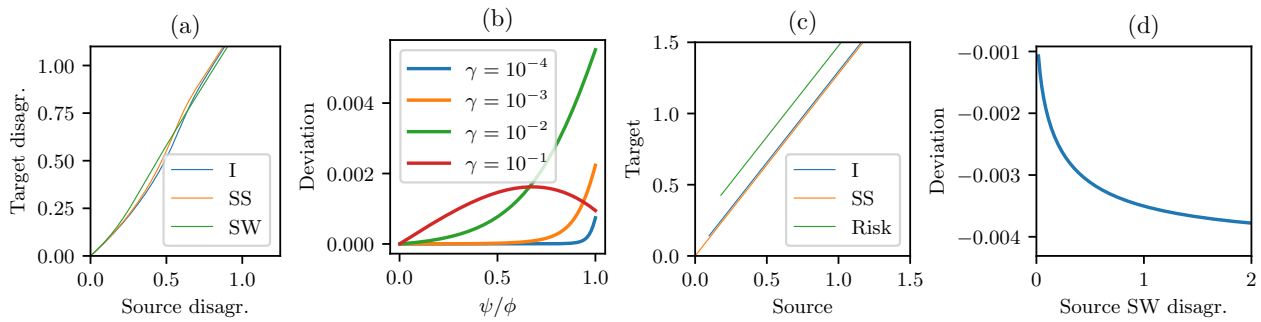

Figure 3. **(a)** Target vs. source I, SS, SW disagreement in the ridgeless and underparametrized regime ($\phi < \psi$). There is no linear trend in this regime. **(b)** Deviation from the line, $\mathrm{Dis}^{\mathrm{t}}_{\mathrm{SS}}(\phi,\psi,\gamma) - a\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SS}}(\phi,\psi,\gamma)$, as a function of ψ for non-zero γ. The deviation becomes larger as γ increases. See Section D.2 for figures for I disagreement and risk. **(c)** Target vs. source lines for I, SS disagreement and risk, in the overparametrized regime $\psi/\phi \in (0,1)$. The lines have *identical* slopes but different *intercepts*. **(d)** Deviation from the line, $\lim_{\gamma\to 0}\mathrm{Dis}^{\mathrm{t}}_{\mathrm{SW}}(\phi,\psi,\gamma) - a\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SW}}(\phi,\psi,\gamma)$, vs. $\lim_{\gamma\to 0}\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SW}}(\phi,\psi,\gamma)$, in the overparametrized regime ($\phi > \psi$). This shows disagreement-on-the-line does not happen for SW disagreement. We use $\phi = 0.5$, $\sigma_\varepsilon^2 = 10^{-4}$, and ReLU activation σ. We set $\mu = 0.4\delta_{(0.1,1)} + 0.6\delta_{(1,0.1)}$ in (a), (b), (d) and $\mu = 0.5\delta_{(4,1)} + 0.5\delta_{(1,4)}$ in (c).

Theorem 4.1 (Exact linear relation)**.** *Define*

$$a = \frac{\rho_{\mathrm{t}}(\omega_{\mathrm{t}} + \mathcal{I}^{\mathrm{t}}_{1,1})}{\rho_{\mathrm{s}}(\omega_{\mathrm{s}} + \mathcal{I}^{\mathrm{s}}_{1,1})}, \quad b_{\mathrm{SS}} = 0,$$
$$b_{\mathrm{I}} = \frac{2\kappa^2(\sigma_\varepsilon^2 + \phi\mathcal{I}^{\mathrm{s}}_{1,2})(\rho_{\mathrm{t}}\mathcal{I}^{\mathrm{t}}_{2,2} - a\rho_{\mathrm{s}}\mathcal{I}^{\mathrm{s}}_{2,2})}{\rho_{\mathrm{s}}(1 - \kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})}, \tag{8}$$

for κ satisfying (7). We fix ϕ and regard the disagreement $\mathrm{Dis}^j_i(\phi,\psi,\gamma)$, $i \in \{\mathrm{I},\mathrm{SS}\}$, $j \in \{\mathrm{s},\mathrm{t}\}$, as a function of ψ. In the overparametrized regime $\phi > \psi$ and for $i \in \{\mathrm{I},\mathrm{SS}\}$,

$$\lim_{\gamma\to 0}\mathrm{Dis}^{\mathrm{t}}_i(\phi,\psi,\gamma) = a\lim_{\gamma\to 0}\mathrm{Dis}^{\mathrm{s}}_i(\phi,\psi,\gamma) + b_i, \tag{9}$$

where the slope a and the intercept b_{I} are independent of ψ.

Recall from (3) and (5) that $\rho_{\mathrm{s}}, \rho_{\mathrm{t}}, \omega_{\mathrm{s}}, \omega_{\mathrm{s}}$ are constants describing non-linearity of the activation σ, and $\mathcal{I}^{\mathrm{s}}_{a,b}, \mathcal{I}^{\mathrm{t}}_{a,b}$ are statistics summarizing spectra of $\Sigma_{\mathrm{s}}, \Sigma_{\mathrm{t}}$. Therefore, the slope a is determined by the property of $\sigma, \Sigma_{\mathrm{s}}, \Sigma_{\mathrm{t}}$. By plugging in sample covariance, we can build an estimate of the slope in finite-sample settings. Also as a sanity check, if we set $\Sigma_{\mathrm{s}} = \Sigma_{\mathrm{t}}$, then we recover $a = 1$ and $b_{\mathrm{I}} = 0$ as there will be no difference between source and target domain.

Remark 4.2. The slope $a = \rho_{\mathrm{t}}(\omega_{\mathrm{t}} + \mathcal{I}^{\mathrm{t}}_{1,1})/\rho_{\mathrm{s}}(\omega_{\mathrm{s}} + \mathcal{I}^{\mathrm{s}}_{1,1})$ is same as the slope from Proposition C.3. This is consistent with the empirical observations from Baek et al. (2022) that the linear trend between ID disagreement and OOD disagreement has the *same slope* as the linear trend between ID risk and OOD risk. However, unlike in Baek et al. (2022), in our case, the intercepts can be different. This can be seen in Figure 1 and Figure 3 **(c)**, and also from (51).

Our analysis provides an explicit formula for the intercepts. Specifically, the intercepts can be numerically computed using equations (8), (51), and Theorem C.1 if σ_ε^2 is known. Note that in the general case of non-linear generative models, σ_ε^2 corresponds to the sum of the noise level and the non-linear component of the data-generating function. By estimating σ_ε^2, we can obtain estimates of the intercepts which can be potentially used for OOD performance estimation.

Ridge. When $\gamma > 0$, the exact linear relation between source disagreement and target disagreement no longer holds in our model. However, it turns out that there is still an *approximate linear relation*, as we show next.

Theorem 4.3 (Approximate linear relation of disagreement)**.** *Let $a, b_{\mathrm{SS}}, b_{\mathrm{I}}$ be defined as in (8). Given $\phi > \psi$, deviation from the line, for I and SS disagreement, is bounded by*

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

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

We see the upper bounds vanish as $\gamma \to 0$, consistent with Theorem 4.1. Also, the upper bound for SS disagreement vanishes as $\psi \to 0$, which is confirmed in Figure 3 **(b)**.

We now present an analog of Theorem 4.3 for prediction error of the random features model. This is a generalization of Proposition C.3, which shows an exact linear relation between risks in the ridgeless and overparametrized regime.

Corollary 4.4 (Approximate linear relation of risk)**.** *Denote prediction risk in the source and target domains by $E_{\mathrm{s}}, E_{\mathrm{t}}$, respectively (see Section C for definitions). Let a, b_{risk} be*

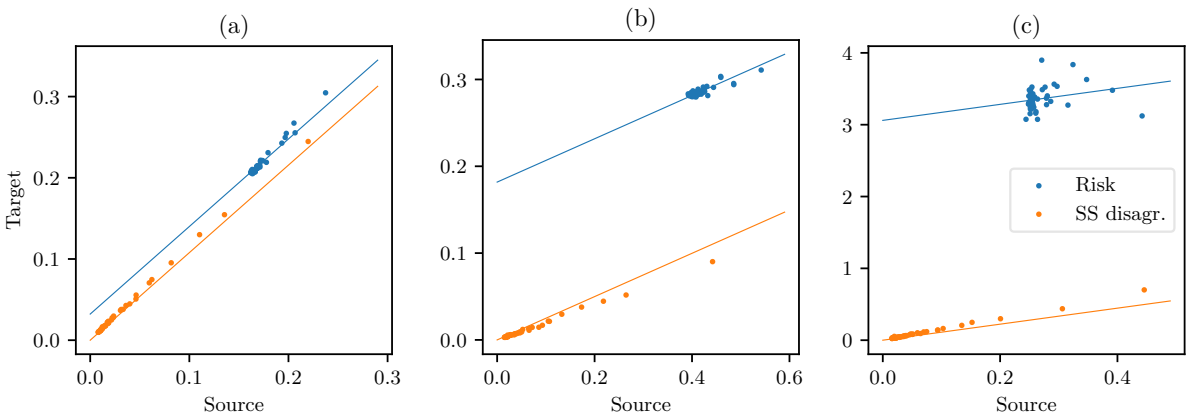

Figure 4. **(a)** CIFAR-10-C-Snow (severity 3) **(b)** Tiny ImageNet-C-Fog (severity 3) **(c)** Camelyon17; For more results, see Section D.3.

defined as in (8) *and* (51). *Given* $\phi > \psi$, *deviation from the line, for risk, is bounded by*

$$|E_{\text{t}} - aE_{\text{s}} - b_{\text{risk}}| \le \frac{C(\gamma + \sqrt{\psi\gamma} + \psi\gamma + \gamma\sqrt{\psi\gamma} + \psi\gamma^2)}{(1 - \psi/\phi + \sqrt{\psi\gamma})^2},$$

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

Theorem 4.3 and Corollary 4.4 together show that the phenomenon we discussed in Remark 4.2 occurs, at least approximately, even when applying ridge regularization.

In the underparametrized case $\psi > \phi$, the self-consistent equation (7) is dependent on ψ, and so is κ. Hence, there is no analog of the linear relation we find in Theorem 4.1 in this regime. Figure 3 **(a)** displays this phenomenon.

4.2. SW disagreement

In Corollary 3.2, unlike I and SS disagreement, SW disagreement contains two linearly independent functions of ψ. Hence, the disagreement-on-the-line phenomenon (9) cannot occur for any choice of slope and intercept independent of ψ. Figure 3 **(a)** and **(d)** confirm the non-linear relation between target vs. source SW disagreement in underparametrized and overparametrized regimes, respectively.

5. Experiments

5.1. Experiments Setup

We conduct experiments on the following datasets. The associated code can be found at `https://github.com/dh7401/RF-disagreement`.

CIFAR-10-C. Hendrycks & Dietterich (2018) introduced a corrupted version of CIFAR-10 (Krizhevsky et al., 2009). We choose two classes and assign the label $y \in \{0, 1\}$ to each. We use CIFAR-10 as the source domain and CIFAR-10-C as the target domain.

Tiny ImageNet-C. Tiny ImageNet (Wu et al., 2017), a smaller version of ImageNet (Deng et al., 2009), consists of natural images of size 64×64 in 200 classes. Tiny ImageNet-C (Hendrycks & Dietterich, 2019) is a corrupted version of Tiny ImageNet. We down-sample images to 32×32 and create two super-classes each consisting of 10 of the original classes. We consider Tiny ImageNet as the source domain and Tiny ImageNet-C as the target domain.

Camelyon17. Camelyon17 (Bandi et al., 2018) consists of tissue slide images collected from five different hospitals, and the task is to identify tumor cells in the images. Koh et al. (2021) proposed a patch-based variant of the task, where the input x is 96×96 image and the label $y \in \{0, 1\}$ indicates whether the central 32×32 contains any tumor tissue. We crop the central 32×32 region and use it as the input in our problem. We use Hospital 0 as the source domain and Hospital 2 as the target domain.

We run random features regression with ReLU activation on these datasets. We use training sample size $n = 1000$, random features dimension $N \in \{3000, 4000, \ldots, 49000\}$, input dimension $d = 3072$, regularization $\gamma = 0$. We test the trained model on the rest of the sample and plot target vs. source SS disagreement and risk. Plots for I and SW disagreements can be found in Section D.4.

We estimate the covariance Σ_{s} and Σ_{t} using the test sample and derive the theoretical slope of target vs. source line predicted by Theorem 4.1 (see Section D.1). Since the limiting spectral distribution of sample covariance is generally different from that of population covariance, we remark that this may lead to a biased estimate of the slope. As the intercept b_{risk} involves the unknown noise level σ_ε^2, it is difficult to make a theoretical prediction on its value. For this reason,

we fit the intercept instead of using its theoretical value.

5.2. Results

While Theorem 4.1 is proved only for Gaussian input and linear generative model, we observe the *disagreement-on-the-line* phenomenon on all three datasets (Figure 4), in which these assumptions are violated.

In this regard, a flurry of recent research (see e.g., Hastie et al. (2022); Hu & Lu (2022); Loureiro et al. (2021); Goldt et al. (2022); Wang et al. (2022); Dudeja et al. (2022); Montanari & Saeed (2022); Pesce et al. (2023)) has proved that findings assuming Gaussian inputs often hold in a much wider range of models. While none of the existing work exactly fits the setting considered in this paper, this gives yet another indication that our theory should remain true more generally. The rigorous characterization of this universality is left for future work.

Also, we find that target vs. source risk does not exhibit a clear linear trend, especially in Tiny ImageNet and Camelyon17. This is because Proposition C.3 does not hold in the case of *concept shift*, i.e., the shift in $\mathbb{P}(y|x)$. However, since disagreement is oblivious to the change of $\mathbb{P}(y|x)$, the *disagreement-on-the-line* is a general phenomenon happening regardless of the type of distribution shift.

6. Conclusion

In this paper, we propose a framework to study various types of disagreement in the random features model. We precisely characterize disagreement in high dimensions and study how disagreement under the source and target domains relate to each other. Our results show that the occurrence of disagreement-on-the-line in the random features model can vary depending on the type of disagreement, regularization, and regime of parameterization. We show that, contrary to the prior observation, the line for disagreement and the line for risk can differ in their intercepts. We run experiments on several real-world datasets and show that the results hold in settings more general than the theoretical setting that we consider.

When the above factors are not properly considered, OOD performance estimation using the disagreement-on-the-line phenomenon can be inaccurate and unreliable. Our findings indicate a potential for further examination of the disagreement-on-the-line principle.

Acknowledgements

The work of Behrad Moniri is supported by The Institute for Learning-enabled Optimization at Scale (TILOS), under award number NSF-CCF-2112665. Donghwan Lee was supported in part by ARO W911NF-20-1-0080, DCIST, Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) #FA9550-20-1-0111 award; Xinmeng Huang was supported in part by the NSF DMS 2046874 (CAREER), NSF CAREER award CIF-1943064.

References

Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, 2020a.

Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, 2020b.

Adlam, B., Levinson, J. A., and Pennington, J. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Anderson, G. W. Convergence of the largest singular value of a polynomial in independent Wigner matrices. *The Annals of Probability*, 41(3B):2103–2181, 2013.

Anderson, G. W. and Zeitouni, O. A CLT for a band matrix model. *Probability Theory and Related Fields*, 134(2): 283–338, 2006.

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

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

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

Bai, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the Camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 38(2):550–560, 2018.

Banna, M., Merlevède, F., and Peligrad, M. On the limiting spectral distribution for a large class of symmetric random matrices with correlated entries. *Stochastic Processes and their Applications*, 125(7):2700–2726, 2015.

Banna, M., Najim, J., and Yao, J. A CLT for linear spectral statistics of large random information-plus-noise matrices. *Stochastic Processes and their Applications*, 130(4):2250–2281, 2020.

Benaych-Georges, F. Rectangular random matrices, related convolution. *Probability Theory and Related Fields*, 144 (3):471–515, 2009.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. In *Proc. Joint European Conf. Mach. Learning and Knowledge Discovery in Databases*, pp. 387–402, 2013.

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

Cheng, X. and Singer, A. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Couillet, R. and Debbah, M. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Couillet, R. and Liao, Z. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

d'Ascoli, S., Gabrié, M., Sagun, L., and Biroli, G. On the interplay between data structure and loss function in classification problems. In *Advances in Neural Information Processing Systems*, 2021.

Deev, A. Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size. In *Sov. Math. Dokl.*, volume 11, pp. 1547–1550, 1970.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Deng, W. and Zheng, L. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Dobriban, E. and Sheng, Y. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

Dobriban, E. and Sheng, Y. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021.

Dobriban, E. and Wager, S. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Dudeja, R., Sen, S., and Lu, Y. M. Spectral universality of regularized linear regression with nearly deterministic sensing matrices. *arXiv preprint arXiv:2208.02753*, 2022.

Dyson, F. J. A Brownian-motion model for the eigenvalues of a random matrix. *Journal of Mathematical Physics*, 3 (6):1191–1198, 1962.

El Karoui, N. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Engel, A. and Van den Broeck, C. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Erdös, L. The matrix Dyson equation and its applications for random matrices. *arXiv preprint arXiv:1903.10060*, 2019.

Erdös, L., Péché, S., Ramírez, J. A., Schlein, B., and Yau, H.-T. Bulk universality for Wigner matrices. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 63(7):895–925, 2010.

Erdös, L., Yau, H.-T., and Yin, J. Bulk universality for generalized Wigner matrices. *Probability Theory and Related Fields*, 154(1):341–407, 2012.

Fan, Z. and Montanari, A. The spectral norm of random inner-product kernel matrices. *Probability Theory and Related Fields*, 173(1):27–85, 2019.

Far, R. R., Oraby, T., Bryc, W., and Speicher, R. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

Far, R. R., Oraby, T., Bryc, W., and Speicher, R. On slow-fading MIMO systems with nonseparable correlation. *IEEE Transactions on Information Theory*, 54(2):544–553, 2008.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*, 2021.

Gaudin, M. Sur la loi limite de l'espacement des valeurs propres d'une matrice ale´ atoire. *Nuclear Physics*, 25: 447–458, 1961.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029–1054, 2021.

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

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., and Schmidt, L. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Györgyi, G. and Tishby, N. Statistical theory of learning a rule. *Neural networks and spin glasses*, pp. 3–36, 1990.

Haagerup, U. and Thorbjørnsen, S. A new application of random matrices: $\mathrm{Ext}(C^*_{\mathrm{red}}(F_2))$ is not a group. *Annals of Mathematics*, 162(2):711–775, 2005.

Haagerup, U., Schultz, H., and Thorbjørnsen, S. A random matrix approach to the lack of projections in $C^*_{\mathrm{red}}(\mathbb{F}_2)$. *Advances in Mathematics*, 204(1):1–83, 2006.

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

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

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Helton, J. W., Far, R. R., and Speicher, R. Operator-valued semicircular elements: solving a quadratic matrix equation with positivity constraints. *International Mathematics Research Notices*, 2007(9), 2007.

Helton, J. W., Mai, T., and Speicher, R. Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., and Lakshminarayanan, B. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020.

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

Hu, H. and Lu, Y. M. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

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

Kirsch, A. and Gal, Y. A note on "assessing generalization of SGD via disagreement". *Transactions on Machine Learning Research*, 2022.

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. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lei, Q., Hu, W., and Lee, J. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, 2021.

Lin, L. and Dobriban, E. What causes the test error? going beyond bias-variance via ANOVA. *Journal of Machine Learning Research*, 22:155–1, 2021.

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

Mehta, M. L. *Random matrices*. Elsevier, 2004.

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

Mel, G. and Pennington, J. Anisotropic random feature regression in high dimensions. In *International Conference on Learning Representations*, 2021.

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

Mingo, J. A. and Speicher, R. *Free probability and random matrices*, volume 35. Springer, 2017.

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

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

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

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Opper, M. Statistical mechanics of learning: Generalization. *The Handbook of Brain Theory and Neural Networks,*, pp. 922–925, 1995.

Opper, M. and Kinzel, W. Statistical mechanics of generalization. In *Models of Neural Networks III*, pp. 151–209. Springer, 1996.

Paul, D. and Aue, A. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

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

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

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2007.

Raudys, Š. On determining training sample size of linear classifier. *Computing Systems (in Russian)*, 28:79–87, 1967.

Raudys, Š. On the amount of a priori information in designing the classification algorithm. *Technical Cybernetics (in Russian)*, 4:168–174, 1972.

Raudys, Š. and Young, D. M. Results in statistical discriminant analysis: A review of the former Soviet Union literature. *Journal of Multivariate Analysis*, 89(1):1–35, 2004.

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

Serdobolskii, V. I. *Multiparametric Statistics*. Elsevier, 2007.

Shlyakhtenko, D. Random Gaussian band matrices and freeness with amalgamation. *International Mathematics Research Notices*, 1996(20):1013–1025, 1996.

Shlyakhtenko, D. Gaussian random band matrices and operator-valued free probability theory. *Banach Center Publications*, 43(1):359–368, 1998.

Speicher, R. *Combinatorial theory of the free product with amalgamation and operator-valued free probability theory*, volume 627. American Mathematical Society, 1998.

Speicher, R. and Vargas, C. Free deterministic equivalents, rectangular random matrix models, and operator-valued free probability theory. *Random Matrices: Theory and Applications*, 1(2), 2012.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tao, T. and Vu, V. Random matrices: universality of local eigenvalue statistics. *Acta mathematica*, 206(1):127–204, 2011.

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

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

Tulino, A. M. and Verdú, S. Random matrix theory and wireless communications. *Communications and Information Theory*, 1(1):1–182, 2004.

Voiculescu, D. Addition of certain non-commuting random variables. *Journal of Functional Analysis*, 66(3):323–346, 1986.

Voiculescu, D. Symmetries of some reduced free product c^*-algebras. In *Operator Algebras and their Connections with Topology and Ergodic Theory: Proceedings of the OATE Conference held in Buşteni, Romania, Aug. 29–Sept. 9, 1983*, pp. 556–588. Springer, 2006.

Volčič, J. Matrix coefficient realization theory of noncommutative rational functions. *Journal of Algebra*, 499: 397–437, 2018.

Wang, T., Zhong, X., and Fan, Z. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

Wigner, E. P. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, pp. 548–564, 1955.

Wu, D. and Xu, J. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.

Wu, J., Zhang, Q., and Xu, G. Tiny ImageNet challenge. *Technical report*, 2017.

Wu, J., Zou, D., Braverman, V., Gu, Q., and Kakade, S. M. The power and limitation of pretraining-finetuning for linear regression under covariate shift. In *Advances in Neural Information Processing Systems*, 2022.

Yao, J., Bai, Z., and Zheng, S. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, 2015.

A. Technical Tools

A.1. Operator-valued Free Probability

Operator-valued free probability (e.g., Speicher (1998); Mingo & Speicher (2017); Helton et al. (2007)) has appeared in various studies of random features models including Adlam et al. (2022); Adlam & Pennington (2020a;b); Mel & Pennington (2021); Ba et al. (2022). Here, we briefly outline the most relevant concepts, which are used in our computation.

Recall that a set $\mathcal{A}$ is an *algebra* (over the field $\mathbb{C}$ of complex numbers) if it is a vector space over $\mathbb{C}$ and is endowed with a *bilinear* multiplication operation denoted by "$\cdot$". Thus, for all $a, b, c \in \mathcal{A}$ we have the distributivity relations $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(b + c) \cdot a = b \cdot a + c \cdot a$; and the relation indicating that multiplication in the algebra is compatible with the usual multiplication over $\mathbb{C}$, namely that for and $x, y \in \mathbb{C}$, $(x \cdot y) \cdot (a \cdot b) = (x \cdot a) \cdot (y \cdot b)$. All algebras we consider will be associative, so that the multiplication operation over the algebra is associative. Further, an algebra is called *unital* if it contains a multiplicative identity element; this is denoted as "1". Often, we drop the "$\cdot$" symbol to denote multiplication (both over the algebra and by scalars), and no confusion may arise.

Definition A.1 (Non-commutative probability space)**.** Let $\mathcal{C}$ be a unital algebra and $\varphi : \mathcal{C} \to \mathbb{C}$ be a linear map such that $\varphi(1) = 1$. We call the pair $(\mathcal{C}, \varphi)$ a *non-commutative probability space*.

Example A.2 (Deterministic matrices). For a matrix $A \in \mathbb{C}^{m\times m}$, we denote its normalized trace by $\overline{\mathrm{tr}}(A) = \frac{1}{m}\sum_{i=1}^{m} A_{ii}$. The pair $(\mathbb{C}^{m\times m}, \overline{\mathrm{tr}})$ is a non-commutative probability space.

Example A.3 (Random matrices). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a (classical) probability space and $L^{-\infty}(\Omega)$ be the set of scalar random variables with all moments finite. The pair $(L^{-\infty}(\Omega)^{m\times m}, \mathbb{E}\overline{\mathrm{tr}})$ is a non-commutative probability space.

Definition A.4 (Operator-valued probability space)**.** Let $\mathcal{A}$ be a unital algebra and consider a unital sub-algebra $\mathcal{B} \subseteq \mathcal{A}$. A linear map $E : \mathcal{A} \to \mathcal{B}$ is a *conditional expectation* if $E(b) = b$ for all $b \in \mathcal{B}$ and $E(b_1 a b_2) = b_1 E(a) b_2$ for all $a \in \mathcal{A}$ and $b_1, b_2 \in \mathcal{B}$. The triple $(\mathcal{A}, E, \mathcal{B})$ is called an *operator-valued probability space*.

The name "conditional expectation" can be understood from the following example.

Example A.5 (Classical conditional expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Then, considering $E = \mathbb{E}[\cdot|\mathcal{G}]$, any unital algebra $\mathcal{A} \subset L^1(\Omega, \mathcal{F}, \mathbb{P})$ and its unital sub-algebra $\mathcal{B} \subset L^1(\Omega, \mathcal{G}, \mathbb{P})$, such that all required integrals in the definition of $E(b_1 a b_2) = b_1 E(a) b_2$ exist for all $a \in \mathcal{A}$ and $b_1, b_2 \in \mathcal{B}$, form an operator-valued probability space $(\mathcal{A}, E, \mathcal{B})$.

Example A.6 (block random matrices). Let $(\mathcal{C}, \varphi) = (L^{-\infty}(\Omega)^{m\times m}, \mathbb{E}\overline{\mathrm{tr}})$ be the non-commutative probability space of random matrices defined in Example A.3. Define $\mathcal{A} = \mathbb{C}^{M\times M} \otimes \mathcal{C}$ and $\mathcal{B} = \mathbb{C}^{M\times M}$. In words, $\mathcal{A}$ is the space of $M \times M$ block matrices with entries in $\mathcal{C}$, and $\mathcal{B}$ is the space of $M \times M$ scalar matrices. Note that $\mathcal{B}$ can be viewed as a unital sub-algebra of $\mathcal{A}$ by the canonical inclusion $\iota : \mathcal{A} \hookrightarrow \mathcal{B}$ defined by

$$\iota(B) = B \otimes 1_{\mathcal{C}}, \tag{10}$$

where $1_{\mathcal{C}}$ is the unity of $\mathcal{C}$ (in this example $1_{\mathcal{C}} = I_m$). We also define the block-wise normalized expected trace $E = \mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}} : \mathcal{A} \to \mathcal{B}$ by

$$E(A) = (\mathbb{E}\overline{\mathrm{tr}} A_{ij})_{1\le i,j\le M}, \quad A = (A_{ij})_{1\le i,j\le M} \in \mathcal{A}. \tag{11}$$

Remark A.7. While we have only discussed squared blocks with identical sizes in Example A.6, it is possible to extend the definition to block matrices with rectangular blocks (Far et al., 2006; 2008; Benaych-Georges, 2009; Speicher & Vargas, 2012). The idea of Benaych-Georges (2009) is to embed each rectangular matrix into a block of a common larger square matrix. For example, if we have rectangular blocks whose dimensions are one of $q_1, \ldots, q_K \in \mathbb{N}$, we consider the space of $(q_1 + \cdots + q_K) \times (q_1 + \cdots + q_K)$ square matrices with a block structure

$$\left[\begin{array}{c|c|c} q_1 \times q_1 & \cdots & q_1 \times q_K \\ \hline \vdots & \ddots & \vdots \\ \hline q_K \times q_1 & \cdots & q_K \times q_K \end{array}\right].$$

Then, we identify a rectangular matrix $C \in \mathbb{C}^{q_i \times q_j}$ with a square matrix $\widetilde{C} \in \mathbb{C}^{(q_1+\cdots+q_K)\times(q_1+\cdots+q_K)}$, having the aforementioned block structure, whose (i, j)-block is C and all other blocks are zero. This identification preserves scalar

multiplication, addition, multiplication, transpose, and trace, in the sense that, for rectangular matrices C, D and a scalar $c \in \mathbb{C}$,

$$c\widetilde{C} = \widetilde{cC}, \quad \widetilde{C} + \widetilde{D} = \widetilde{C+D} \quad \text{if } C \text{ and } D \text{ have same shape}, \quad (\widetilde{C})^\top = \widetilde{C^\top},$$
$$\widetilde{C}\widetilde{D} = \begin{cases} \widetilde{CD} & \text{if } C \text{ and } D \text{ are conformable,} \\ 0 & \text{otherwise,} \end{cases} \qquad \mathrm{tr}(\widetilde{C}) = \begin{cases} \mathrm{tr}(C) & \text{if } C \text{ is a square matrix,} \\ 0 & \text{otherwise.} \end{cases}$$

Through this identification, the space of rectangular matrices (with finitely many different dimension types) can be also understood as an algebra over $\mathbb{C}$. Further, by replacing $\mathcal{C}$ in Example A.6 with the space of rectangular random matrices, we can define the space of block random matrices with rectangular blocks. The space of block random matrices with rectangular blocks, equipped with the block-wised expected trace, will be the operator-valued probability space we consider in our proof.

Definition A.8 (Operator-valued Cauchy transform)**.** Let $(\mathcal{A}, E, \mathcal{B})$ be an operator-valued probability space. For $a \in \mathcal{A}$, define its operator-valued Cauchy transform $\mathcal{G}_a : \mathcal{B} \setminus \{a\} \to \mathcal{B}$ by

$$\mathcal{G}_a(b) = E[(b-a)^{-1}].$$

Definition A.9 (Operator-valued freeness)**.** Let $(\mathcal{A}, E, \mathcal{B})$ be an operator-valued probability space and $(\mathcal{A}_i)_{i\in I}$ be a family of sub-algebras of $\mathcal{A}$ which contain $\mathcal{B}$. The sub-algebras $\mathcal{A}_i$ are *freely independent* over $\mathcal{B}$, if $E[a_1 \cdots a_n] = 0$ whenever $E[a_1] = \cdots = E[a_n] = 0$ and $a_i \in \mathcal{A}_{j(i)}$ for all $i \in [n]$ with $j(1) \neq \cdots \neq j(n)$. Variables $a_1, \ldots, a_n \in \mathcal{A}$ are freely independent over $\mathcal{B}$ if the sub-algebras generated by a_i and $\mathcal{B}$ are freely independent over $\mathcal{B}$.

Another important transform, introduced in Voiculescu (1986; 2006), is the R-transform. It enables the characterization of the spectrum of a sum of asymptotically freely independent random matrices. It was generalized to operator-valued probability spaces in Shlyakhtenko (1996); Mingo & Speicher (2017). The definition of operator-valued R-transform can be found in Definition 10, Chapter 9 of Mingo & Speicher (2017). Our work does not directly require the definition of R-transforms, and instead uses the following property.

Proposition A.10 (Subordination property, (9.21) of Mingo & Speicher (2017))**.** *Let $(\mathcal{A}, E, \mathcal{B})$ be an operator-valued probability space. If $x, y \in \mathcal{A}$ are freely independent over $\mathcal{B}$, then*

$$\mathcal{G}_{x+y}(b) = \mathcal{G}_x[b - \mathcal{R}_y(\mathcal{G}_{x+y}(b))] \tag{12}$$

for all $b \in \mathcal{B}$, where $\mathcal{R}_y$ is the operator-valued R-transform of y.

A.2. Limiting R-transform of Gaussian Block Matrices

Shlyakhtenko (1996; 1998) proposed using operator-valued free probability to study spectra of Gaussian block matrices. Their insight was that operator-valued free independence among Gaussian block matrices is guaranteed for general covariance structure, whereas scalar-valued freeness among them only holds in special cases. Later Far et al. (2006; 2008); Anderson & Zeitouni (2006) revisited this idea. We present a theorem of Far et al. (2008), which characterizes limiting R-transform of Gaussian block matrices with rectangular blocks.

Theorem A.11 (Theorem 5 of Far et al. (2008))**.** *For $m = m_1 + \cdots + m_M$, let $A = (A_{ij})_{1\le i,j\le M} \in \mathbb{R}^{m\times m}$ be an $M \times M$ block random matrix whose block A_{ij} is a $m_i \times m_j$ random matrix with i.i.d. $\mathsf{N}(0, c_{ij}^2/m)$ entries. Define the covariance function $\sigma(i,j;k,l)$ to be $c_{ij}c_{kl}$ if $A_{ij}/c_{ij} = A_{kl}^\top/c_{kl}$ and 0 otherwise. We assume the proportional limit where $m_1, \ldots, m_M \to \infty$ with $m_i/m \to \alpha_i \in (0,\infty)$, $i = 1, \ldots, M$. Then, the limiting R-transform of A can be expressed as*

$$[\mathcal{R}_A(D)]_{ij} = \sum_{1\le k,l\le M} \sigma(i,k;l,j)\alpha_k D_{kl}, \tag{13}$$

for any $D \in \mathbb{R}^{M\times M}$.

We remark the above statement should be understood in the space of block random matrices with rectangular blocks we discussed in Remark A.7. Also, the original statement used a different terminology "covariance mapping", but it is identical to the R-transform of A (see discussion in Mingo & Speicher (2017) p.242 and Far et al. (2006) p.24)

A.3. Centering Random Features

We first argue that the random features F, f can be centered without changing the asymptotics of disagreement. This centering argument became a standard technique after it was introduced in Mei & Montanari (2022) (Section 10.4). More generally, centering arguments are standard in random matrix theory (see e.g., Bai & Silverstein (2010)). For a standard Gaussian random variable $Z \sim \mathsf{N}(0,1)$, define centered random features by

$$\bar{F} = F - \mathbb{E}\sigma(\sqrt{m_{\mathrm{s}}}Z), \quad \bar{f} = f - \mathbb{E}\sigma(\sqrt{m_j}Z),$$

where $j \in \{\mathrm{s}, \mathrm{t}\}$ is the domain that input x comes from. Subtracting a scalar from a matrix/vector should be understood entry-wise. The following lemma states that model prediction obtained from these centered random features is close to the original prediction $\hat{y}(x)$ with high probability.

Lemma A.12. *Define centered model prediction by*

$$\bar{\hat{y}}(x) = Y^\top \left(\frac{1}{N}\bar{F}^\top \bar{F} + \gamma I_n\right)^{-1} \left(\frac{1}{N}\bar{F}^\top \bar{f}\right).$$

There exist constants $c_1, c_2, c_3, c_4 > 0$ such that

$$|\bar{\hat{y}}(x) - \hat{y}(x)| \le c_1 d^{-c_2}$$

with probability at least $1 - c_3 d^{-c_4}$.

This lemma is a consequence of Lemma I.7 and Lemma I.8 of Tripuraneni et al. (2021). Since we consider the limit $n, d, N \to \infty$, disagreement $\mathrm{Dis}_i(\phi, \psi, \gamma), i \in \{\mathrm{I}, \mathrm{SS}, \mathrm{SW}\}$ are invariant to the centering. We also remark that the non-linearity constants defined in (3) are also unchanged after this centering. For these reasons, perhaps with a slight abuse of notation, we assume F and f are centered from now on.

A.4. Gaussian Equivalence

For domain $j \in \{\mathrm{s}, \mathrm{t}\}$ that input x is drawn from, we consider the following *noisy linear* random features

$$\tilde{F} = \sqrt{\frac{\rho_{\mathrm{s}}}{d}}WX + \sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta, \quad \tilde{f} = \sqrt{\frac{\rho_j}{d}}Wx + \sqrt{\rho_j\omega_j}\theta, \tag{14}$$

where $\Theta \in \mathbb{R}^{N\times n}$ and $\theta \in \mathbb{R}^N$ have i.i.d. standard Gaussian entries independent from all other Gaussian matrices. The coefficients above are chosen so that the first and second moment of $\tilde{F}$ and $\tilde{f}$ match those of F and f, respectively. We call $\tilde{F}, \tilde{f}$ the *Gaussian equivalent* of F, f as we claim the following.

Claim A.13 (Gaussian equivalence)**.** The asymptotic limit (Condition 2.2) of the disagreement (Definition 2.1) of the random features model (2) is invariant to the substitution $F, f \to \tilde{F}, \tilde{f}$.

This idea was introduced in the context of random kernel matrices (El Karoui, 2010; Cheng & Singer, 2013; Fan & Montanari, 2019) and has been repeatedly used in recent studies of random feature models. Mei & Montanari (2022) proved the Gaussian equivalence for random weights uniformly distributed on a sphere. Montanari et al. (2019) conjectured that the same holds for classification. Adlam & Pennington (2020a;b); Tripuraneni et al. (2021) derived several asymptotic properties of random features models building on the Gaussian equivalence conjecture. Goldt et al. (2022) provided theoretical and numerical evidence suggesting that the Gaussian equivalence holds for a wide class of models including random features models. Mel & Pennington (2021); d'Ascoli et al. (2021); Loureiro et al. (2021) conjectured the Gaussian equivalence for anisotropic inputs. Hassani & Javanmard (2022) showed the Gaussian equivalence holds for the adversarial risk of adversarially trained random features models. Hu & Lu (2022) showed the conjecture for isotropic Gaussian inputs, under mild technical conditions. Montanari & Saeed (2022) generalized this by removing the isotropic condition and relaxing the Gaussian input assumption.

More generally, the phenomenon that eigenvalue statistics in the bulk spectrum of a random matrix do not depend on the specific law of the matrix entries is referred to as "bulk universality" (Wigner, 1955; Gaudin, 1961; Mehta, 2004; Dyson, 1962) and has been a central subject in the random matrix theory literature (Erdös et al., 2010; 2012; El Karoui, 2010; Tao & Vu, 2011).

It is known that local spectral laws of correlated random hermitian matrices can be fully determined by their first and second moments, through the matrix Dyson equation (Erdös, 2019). Also, Banna et al. (2015; 2020) showed that spectral distributions of correlated symmetric random matrices and sample covariance matrices can be characterized by Gaussian matrices with identical correlation structures. However, these results do not directly imply Claim A.13 since we do not study the spectral properties of F, f on their own.

A.5. Linear Pencils

After applying the Gaussian equivalence (14), each of the quantities that we study becomes an expected trace of a rational function of random matrices. To analyze this, we use the *linear pencil* method (Haagerup & Thorbjørnsen, 2005; Haagerup et al., 2006; Anderson, 2013; Helton et al., 2018), in which we build a large block matrix whose blocks are linear functions of variables and one of the blocks of its inverse is the desired rational function. Then, operator-valued free probability can be used to extract block-wise spectral properties of the inverse. For example, if we want to compute $\mathbb{E}\,\mathrm{tr}[(\frac{X^\top X}{d}+\gamma I_n)^{-1}]$ for $X \in \mathbb{R}^{d\times n}$, we consider

$$\begin{bmatrix} I_n & -\frac{X^\top}{\sqrt{\gamma d}} \\ \frac{X}{\sqrt{\gamma d}} & I_d \end{bmatrix},$$

inverse has as its (1, 1)-block $\gamma(\frac{X^\top X}{d}+\gamma I_n)^{-1}$. Block matrices for more complicated rational functions can be constructed using the following proposition.

Proposition A.14 (Algorithm 4.3 of Helton et al. (2018))**.** *Let $x_1, \ldots, x_g$ be elements of an algebra $\mathcal{A}$ over a field $\mathbb{K}$. For an $m \times m$ matrix Q and vectors $u, v \in \mathbb{K}^m$, a triple (u, Q, v) is called a linear pencil of a rational function $r \in \mathbb{K}(x_1, \ldots, x_g)$ if each entry of Q is a $\mathbb{K}$-affine function of $x_1, \ldots, x_g$ and $r = -u^\top Q^{-1} v$. The following holds.*

1. *(Addition) If (u_1, Q_1, v_1) and (u_2, Q_2, v_2) are linear pencils of r_1 and r_2, respectively, then*

$$\left(\begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \begin{bmatrix} Q_1 & 0_{m\times m} \\ 0_{m\times m} & Q_2 \end{bmatrix}, \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}\right)$$

   is a linear pencil of $r_1 + r_2$.

2. *(Multiplication) If (u_1, Q_1, v_1) and (u_2, Q_2, v_2) are linear pencils of r_1 and r_2, respectively, then*

$$\left(\begin{bmatrix} 0_m \\ u_1 \end{bmatrix}, \begin{bmatrix} x_g v_1 u_2^\top & Q_1 \\ Q_2 & 0_{m\times m} \end{bmatrix}, \begin{bmatrix} 0_m \\ v_2 \end{bmatrix}\right)$$

   is a linear pencil of $r_1 x_g r_2$.

3. *(Inverse) If (u, Q, v) is a linear pencil of r, then*

$$\left(\begin{bmatrix} 1 \\ 0_m \end{bmatrix}, \begin{bmatrix} 0 & u^\top \\ v & -Q^{-1} \end{bmatrix}, \begin{bmatrix} 1 \\ 0_m \end{bmatrix}\right)$$

   is a linear pencil of r^{-1}.

In this language, the example before the algorithm can be interpreted in the space we consider in Remark A.7 as $r = -\gamma(\frac{X^\top X}{d}+\gamma I_n)^{-1}$ being a rational function of X and $X^\top$, and

$$\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} I_n & -\frac{X^\top}{\sqrt{\gamma d}} \\ \frac{X}{\sqrt{\gamma d}} & I_d \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \tag{15}$$

being a linear pencil of r.

In principle, repeated application of the above rules to basic building blocks such as (15) can produce a linear pencil for any rational function of given random matrices. For example, consider $X_1, X_2 \in \mathbb{R}^{d\times n}, \Sigma \in \mathbb{R}^{d\times d}$ and their transpose as

elements of the algebra over $\mathbb{R}$ we discussed in Remark A.7. Then,

$$\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_n & -\frac{X_1^\top}{\sqrt{\gamma d}} & -\frac{\Sigma}{\gamma^2} & \cdot \\ \frac{X_1}{\sqrt{\gamma d}} & I_d & \cdot & \cdot \\ \cdot & \cdot & I_n & -\frac{X_2^\top}{\sqrt{\gamma d}} \\ \cdot & \cdot & \frac{X_2}{\sqrt{\gamma d}} & I_d \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \right)$$

is a linear pencil of $r' = -(\frac{X_1^\top X_1}{d} + \gamma I_n)^{-1}\Sigma(\frac{X_2^\top X_2}{d} + \gamma I_n)^{-1}$. Here, we denote zero blocks by dots. This can be seen by applying the multiplication rule to two copies of (15) and $x_g = \Sigma$, and then switching the first and the second pairs of columns.

However, constructing a suitably small linear pencil is a non-trivial problem of independent interest (see discussions on reductions of linear pencils in e.g., Volčič (2018); Helton et al. (2018) and references therein). This is one of the challenges we need to overcome in our proofs.

B. Proofs

B.1. Proof of Theorem 3.1

Starting from this section, we omit the high-dimensional limit signs $\lim_{n,d,N\to\infty}$ (Condition 2.2) for a simpler presentation. However, every expectation appearing in the derivation should be understood as its high-dimensional limit.

For $j \in \{\mathrm{s}, \mathrm{t}\}$, independent disagreement satisfies

$$\begin{aligned}
&\mathrm{Dis}_{\mathrm{I}}^j(\phi,\psi,\gamma) = \mathbb{E}[(\hat{y}_{W_1,X_1,Y_1}(x) - \hat{y}_{W_2,X_2,Y_2}(x))^2] \\
&= \mathbb{E}[(\hat{y}_{W_1,X_1,Y_1}(x) - \mathbb{E}_{W_1,X_1,Y_1}[\hat{y}_{W_1,X_1,Y_1}(x)] + \mathbb{E}_{W_2,X_2,Y_2}[\hat{y}_{W_2,X_2,Y_2}(x)] - \hat{y}_{W_2,X_2,Y_2}(x))^2] \\
&= \mathbb{E}_{\beta,x\sim\mathcal{D}_j}[(\hat{y}_{W_1,X_1,Y_1}(x) - \mathbb{E}_{W_1,X_1,Y_1}[\hat{y}_{W_1,X_1,Y_1}(x)])^2] + \mathbb{E}_{\beta,x\sim\mathcal{D}_j}[(\hat{y}_{W_2,X_2,Y_2}(x) - \mathbb{E}_{W_2,X_2,Y_2}[\hat{y}_{W_2,X_2,Y_2}(x)])^2] \\
&= \mathbb{E}_{\beta,x\sim\mathcal{D}_j}\mathbb{V}_{W_1,X_1,Y_1}(\hat{y}_{W_1,X_1,Y_1}(x)) + \mathbb{E}_{\beta,x\sim\mathcal{D}_j}\mathbb{V}_{W_2,X_2,Y_2}(\hat{y}_{W_2,X_2,Y_2}(x)) = 2V_j.
\end{aligned}$$

Plugging in the variance V_j given in Theorem C.1, we obtain the formula for $\mathrm{Dis}_{\mathrm{I}}^j(\phi,\psi,\gamma)$.

B.1.1. Decomposition of $\mathrm{Dis}_{\mathrm{SS}}^j(\phi,\psi,\gamma)$

Writing $F_i = \sigma(W_i X/\sqrt{d})$, $f_i = \sigma(W_i x/\sqrt{d})$, $K_i = \frac{1}{N}F_i^\top F_i + \gamma I_n$ for $i \in \{1,2\}$, we can write shared-sample disagreement as

$$\begin{aligned}
\mathrm{Dis}_{\mathrm{SS}}^j(\phi,\psi,\gamma) &= \frac{1}{N^2}\mathbb{E}[(Y^\top K_1^{-1}F_1^\top f_1 - Y^\top K_2^{-1}F_2^\top f_2)^2] \\
&= \frac{2}{N^2}\mathbb{E}[f_1^\top F_1 K_1^{-1} Y Y^\top K_1^{-1} F_1^\top f_1] - \frac{2}{N^2}\mathbb{E}[f_2^\top F_2 K_2^{-1} Y Y^\top K_1^{-1} F_1^\top f_1] \\
&= D_1 - D_2. \qquad (16)
\end{aligned}$$

The term D_1 was computed in (A268), (A279), (A462), (A546) of Tripuraneni et al. (2021) as

$$D_1 = 2V_j + \frac{2\rho_j\kappa^2}{\rho_{\mathrm{s}}\phi}\mathcal{I}_{3,2}^j. \qquad (17)$$

Plugging in $Y = X^\top\beta/\sqrt{d} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} = (\varepsilon_1,\dots,\varepsilon_n)^\top \in \mathbb{R}^n$, the term D_2 becomes

$$\begin{aligned}
D_2 =& \frac{2}{dN^2}\mathbb{E}_{W_i,X}\,\mathrm{tr}[K_2^{-1}X^\top\mathbb{E}_\beta[\beta\beta^\top]XK_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1f_2^\top]F_2] \\
&+ \frac{4}{\sqrt{d}N^2}\mathbb{E}_{W_i,X}[K_2^{-1}X^\top\mathbb{E}_{\beta,\boldsymbol{\varepsilon}}[\beta\boldsymbol{\varepsilon}^\top]K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1f_2^\top]F_2] \\
&+ \frac{2}{N^2}\mathbb{E}_{W_i,X}\,\mathrm{tr}[K_2^{-1}\mathbb{E}_{\boldsymbol{\varepsilon}}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top]K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1f_2^\top]F_2] \\
=& \frac{2}{dN^2}\mathbb{E}_{W_i,X}\,\mathrm{tr}[K_2^{-1}X^\top XK_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1f_2^\top]F_2] + \frac{2\sigma_\varepsilon^2}{N^2}\mathbb{E}_{W_i,X}\,\mathrm{tr}[K_2^{-1}K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1f_2^\top]F_2].
\end{aligned}$$

From the Gaussian equivalence (14), we have

$$\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[f_1 f_2^\top] = \frac{\rho_j}{d} W_1 \Sigma_j W_2^\top.$$

Therefore,

$$\begin{aligned} D_2 &= \frac{2\rho_j}{d^2N^2}\mathbb{E}_{W_i,X}\operatorname{tr}[W_1\Sigma_j W_2^\top F_2 K_2^{-1} X^\top X K_1^{-1} F_1^\top] + \frac{2\sigma_\varepsilon^2\rho_j}{dN^2}\mathbb{E}_{W_i,X}\operatorname{tr}[K_1^{-1}F_1^\top W_1\Sigma_j W_2^\top F_2 K_2^{-1}] \\ &= D_{21} + D_{22}. \end{aligned} \tag{18}$$

We can write $X = \Sigma_{\mathrm{s}}^{\frac{1}{2}} Z$ for $Z \in \mathbb{R}^{d\times n}$ with i.i.d. standard Gaussian entries. Thus,

$$\begin{aligned} D_{21} &= \frac{2\rho_j}{d^2N^2}\mathbb{E}_{W_i,Z}\operatorname{tr}[W_1\Sigma_j W_2^\top F_2 K_2^{-1} Z^\top \Sigma_{\mathrm{s}} Z K_1^{-1} F_1^\top], \\ D_{22} &= \frac{2\sigma_\varepsilon^2\rho_j}{dN^2}\mathbb{E}_{W_i,Z}\operatorname{tr}[K_1^{-1}F_1^\top W_1\Sigma_j W_2^\top F_2 K_2^{-1}]. \end{aligned}$$

Now, we use the linear pencil method (Helton et al., 2018) to build a block matrix such that (1) each block is either deterministic or a constant multiple of Z, W_i, Θ_i and (2) D_{21} or D_{22} appears as a trace of a block of its inverse. Then, we compute the operator-valued Cauchy transform of the block matrix and extract D_{21} and D_{22} from the result.

B.1.2. Preliminary computations

We present some preliminary computations that will be used in later sections. We will also use the linear pencil Q^0 as a building block when constructing other linear pencils. Most of the computations here are adopted from Section A.9.6.1 of Tripuraneni et al. (2021). For clarity and to be self-contained, we provide our own version of the same result updated in some minor ways.

Using W, Z and other notations from Section 2 and Θ from (14), let

$$Q^0 = \begin{bmatrix} I_n & \frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\mathrm{s}}}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot \\ -\frac{\Theta\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\mathrm{s}}}W}{\sqrt{N}} & \cdot \\ \cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot \\ \cdot & -\frac{W^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}} \\ -\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d \end{bmatrix}.$$

Recall from Example A.2 that we denote the normalized trace of a matrix A by $\overline{\operatorname{tr}}(A)$. Define the block-wise normalized expected trace of $(Q^0)^{-1}$ by $G^0 = (\mathrm{id}\otimes\mathbb{E}\overline{\operatorname{tr}})((Q^0)^{-1})$. From block matrix inversion, we see

$$G^0_{1,1} = \gamma\,\mathbb{E}\overline{\operatorname{tr}}(K^{-1}), \quad G^0_{3,6} = \frac{\gamma\sqrt{\rho_{\mathrm{s}}}\,\mathbb{E}\overline{\operatorname{tr}}[\Sigma_{\mathrm{s}} W^\top \hat{K}^{-1} W]}{N}, \quad G^0_{5,4} = -\frac{\sqrt{\rho_{\mathrm{s}}}\,\mathbb{E}\overline{\operatorname{tr}}[\Sigma_{\mathrm{s}} Z K^{-1} Z^\top]}{d}, \tag{19}$$

in which $\hat{K} = \frac{1}{N}FF^\top + \gamma I_N$. We augment the matrix Q^0 to form the symmetric matrix $\bar{Q}^0$ as

$$\bar{Q}^0 = \begin{bmatrix} \cdot & (Q^0)^\top \\ Q^0 & \cdot \end{bmatrix}.$$

This matrix can be written as

$$\begin{aligned} \bar{Q}^0 &= \bar{Z}^0 - \bar{Q}^0_{W,Z,\Theta} - \bar{Q}^0_{\Sigma} \\ &= \begin{bmatrix} \cdot & I_{n+4d+N} \\ I_{n+4d+N} & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & (Q^0_{W,Z,\Theta})^\top \\ Q^0_{W,Z,\Theta} & \cdot \end{bmatrix} - \begin{bmatrix} 0 & (Q^0_{\Sigma})^\top \\ Q^0_{\Sigma} & \cdot \end{bmatrix}, \end{aligned}$$

with

$$Q^0_{W,Z,\Theta} = \begin{bmatrix} \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot \\ \frac{\Theta\sqrt{\rho_s\omega_s}}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W}{\sqrt{N}} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \frac{W^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad \text{and} \quad Q^0_\Sigma = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Defining $\bar{G}^0$ as below, we have

$$\begin{aligned} \bar{G}^0 &= \begin{bmatrix} \cdot & G^0 \\ (G^0)^\top & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((Q^0)^{-1}) \\ (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(((Q^0)^\top)^{-1}) & \cdot \end{bmatrix} \\ &= (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}}) \begin{bmatrix} \cdot & (Q^0)^{-1} \\ ((Q^0)^\top)^{-1} & \cdot \end{bmatrix} = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((\bar{Q}^0)^{-1}). \end{aligned}$$

Thus, $\bar{G}^0$ can be viewed as the operator-valued Cauchy transform of $\bar{Q}^0_{W,Z,\Theta} + \bar{Q}^0_\Sigma$ (in the space we consider in Remark A.7),

$$\bar{G}^0 = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^0 - \bar{Q}^0_{W,Z,\Theta} - \bar{Q}^0_\Sigma)^{-1} = \mathcal{G}_{\bar{Q}^0_{W,Z,\Theta}+\bar{Q}^0_\Sigma}(\bar{Z}^0).$$

Here, we implicitly used the canonical inclusion defined in (10) to write

$$\bar{Z}^0 = \begin{bmatrix} \cdot & I_6 \\ I_6 & \cdot \end{bmatrix}.$$

Since $\bar{Q}^0_\Sigma$ is deterministic, the matrices $\bar{Q}^0_{W,Z,\Theta}$ and $\bar{Q}^0_\Sigma$ are asymptotically freely independent according to Definition A.9. Hence by the subordination formula (12),

$$\bar{G}^0 = \mathcal{G}_{\bar{Q}^0_\Sigma}(\bar{Z}^0 - \mathcal{R}_{\bar{Q}^0_{W,Z,\Theta}}(\bar{G}^0)) = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^0 - \mathcal{R}_{\bar{Q}^0_{W,Z,\Theta}}(\bar{G}^0) - \bar{Q}^0_\Sigma)^{-1}. \tag{20}$$

Since $\bar{Q}^0_{W,Z,\Theta}$ consists of i.i.d. Gaussian blocks, we use (13) to find the R-transform $\mathcal{R}_{\bar{Q}^0_{W,Z,\Theta}}(\bar{G}^0)$ of the form

$$\mathcal{R}_{\bar{Q}^0_{W,Z,\Theta}}(\bar{G}^0) = \begin{bmatrix} \cdot & (R^0)^\top \\ R^0 & \cdot \end{bmatrix}.$$

For example, to find $R^0_{1,1}$, we look for a block in the first row of $\bar{Q}^0_{W,Z,\Theta}$ and a block in the first column of $\bar{Q}^0_{W,Z,\Theta}$ such that they are transpose to each other up to a constant factor. There are two such pairs, ((1, 2)-block, (2, 1)-block) and ((1, 3)-block, (6, 1)-block). Therefore, the equation (13) gives

$$R^0_{1,1} = -\frac{\rho_s\omega_s}{\gamma}G^0_{2,2} - \frac{\sqrt{\rho_s}}{\gamma}G^0_{3,6}.$$

Repeating the same procedure, the non-zero blocks of R^0 are

$$\begin{aligned} &R^0_{1,1} = -\frac{\rho_s\omega_s}{\gamma}G^0_{2,2} - \frac{\sqrt{\rho_s}}{\gamma}G^0_{3,6}, \quad R^0_{2,2} = -\frac{\rho_s\omega_s\psi}{\gamma\phi}G^0_{1,1} + \sqrt{\rho_s}\psi G^0_{5,4}, \\ &R^0_{4,5} = \sqrt{\rho_s}G^0_{2,2}, \quad R^0_{6,3} = -\frac{\sqrt{\rho_s}G^0_{1,1}}{\gamma\phi}. \end{aligned}$$

Plugging this into equation (20), we obtain self-consistent equations for G^1. For example,

$$\begin{aligned} G^0_{3,6} &= \mathbb{E}\overline{\mathrm{tr}}[(I_{n+4d+N} - R^0 - Q^0_\Sigma)]_{3,6} = \mathbb{E}\overline{\mathrm{tr}}\left[\gamma\sqrt{\rho_s}\phi G^0_{2,2}\Sigma_s(\gamma\phi I_d + \rho_s G^0_{1,1}G^0_{2,2}\Sigma_s)^{-1}\right] \\ &= \mathbb{E}_\mu\left[\frac{\lambda^s\gamma\sqrt{\rho_s}\phi G^0_{2,2}}{\gamma\phi + \lambda^s\rho_s G^0_{1,1}G^0_{2,2}}\right]. \end{aligned}$$

Similarly,

$$G^0_{1,1} = \frac{\gamma}{\gamma + \rho_{\rm s}\omega_{\rm s}G^0_{2,2} + \sqrt{\rho_{\rm s}}G^0_{3,6}}, \quad G^0_{2,2} = \frac{\gamma\phi}{\gamma\phi + \rho_{\rm s}\omega_{\rm s}\psi G^0_{1,1} - \gamma\sqrt{\rho_{\rm s}}\psi\phi G^0_{5,4}}$$
$$G^0_{3,6} = \mathbb{E}_\mu\left[\frac{\lambda^{\rm s}\gamma\sqrt{\rho_{\rm s}}\phi G^0_{2,2}}{\gamma\phi + \lambda^{\rm s}\rho_{\rm s}G^0_{1,1}G^0_{2,2}}\right], \quad G^0_{5,4} = -\mathbb{E}_\mu\left[\frac{\lambda^{\rm s}\sqrt{\rho_{\rm s}}G^0_{1,1}}{\gamma\phi + \lambda^{\rm s}\rho_{\rm s}G^0_{1,1}G^0_{2,2}}\right].$$

Now, by eliminating $G^0_{3,6}, G^0_{5,4}$ and expressing in terms of κ, τ, and $\bar\tau$ defined in (4) and (6), we can show that $\mathbb{E}\overline{\rm tr}(K^{-1}) = \frac{G^0_{1,1}}{\gamma} = \tau$, and $\mathbb{E}\overline{\rm tr}(\hat K^{-1}) = \frac{G^0_{2,2}}{\gamma} = \bar\tau$. Thus, using equation (5) we have

$$G^0_{1,1} = \gamma\tau, \quad G^0_{2,2} = \gamma\bar\tau, \quad G^0_{3,6} = \gamma\sqrt{\rho_{\rm s}}\bar\tau\mathcal{I}^{\rm s}_{1,1}, \quad G^0_{5,4} = -\frac{\sqrt{\rho_{\rm s}}\tau\mathcal{I}^{\rm s}_{1,1}}{\phi}. \tag{21}$$

B.1.3. Computation of D_{21}

Define Q^1 by

$$Q^1 = \begin{bmatrix}
I_n & \frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_2^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\rm s}}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_2}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}}W_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac12} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -\frac{W_2^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\Sigma_{\rm s}^{\frac12}}{\sqrt{\rho_{\rm s}}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac12} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_n & \frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_1^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\rm s}}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_1}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}}W_1}{\sqrt{N}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac12} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{W_1^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac12}\frac{\Sigma_j}{\sqrt{\rho_{\rm s}}} & \cdot & \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\frac{W_2^\top}{\sqrt{N}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_N
\end{bmatrix}$$

Define the block-wise normalized expected trace of $(Q^1)^{-1}$ by $G^1 = (\mathrm{id}\otimes\mathbb{E}\overline{\rm tr})((Q^1)^{-1})$. Then, by block matrix inversion we have

$$G^1_{2,14} = \frac{\psi}{d^2N^2}\mathbb{E}\,{\rm tr}[W_1\Sigma_jW_2^\top F_2K_2^{-1}Z^\top\Sigma_{\rm s}ZK_1^{-1}F_1^\top] = \frac{\psi}{2\rho_j}D_{21}.$$

We augment Q^1 to the symmetric matrix $\bar Q^1$ as

$$\bar Q^1 = \begin{bmatrix} \cdot & (Q^1)^\top \\ Q^1 & \cdot \end{bmatrix}$$

and write

$$\begin{aligned}\bar Q^1 &= \bar Z^1 - \bar Q^1_{W,Z,\Theta} - \bar Q^1_\Sigma \\ &= \begin{bmatrix} \cdot & I_{2n+9d+3N} \\ I_{2n+9d+3N} & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & (Q^1_{W,Z,\Theta})^\top \\ Q^1_{W,Z,\Theta} & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & (Q^1_\Sigma)^\top \\ Q^1_\Sigma & \cdot \end{bmatrix},\end{aligned}$$

where

$$Q^1_{W,Z,\Theta} = \begin{bmatrix} \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_2^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\sqrt{\rho_s\omega_s}\Theta_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \frac{W_2^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_1^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s\omega_s}\Theta_1}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W_1}{\sqrt{N}} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W_1^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W_2^\top}{\sqrt{N}} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

and

$$Q^1_\Sigma = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\Sigma_s^{\frac{1}{2}}}{\sqrt{\rho_s}} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & -\frac{\Sigma_j}{\sqrt{\rho_s}} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Then defining $\bar{G}^1$ below,

$$\begin{aligned} \bar{G}^1 &= \begin{bmatrix} \cdot & G^1 \\ (G^1)^\top & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((Q^1)^{-1}) \\ (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(((Q^1)^\top)^{-1}) & \cdot \end{bmatrix} \\ &= (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}}) \begin{bmatrix} \cdot & (Q^1)^{-1} \\ ((Q^1)^\top)^{-1} & \cdot \end{bmatrix} = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((\bar{Q}^1)^{-1}) \end{aligned}$$

can be viewed as the operator-valued Cauchy transform of $\bar{Q}^1_{W,Z,\Theta} + \bar{Q}^1_\Sigma$ (in the space we consider in Remark A.7), i.e.,

$$\bar{G}^1 = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^1 - \bar{Q}^1_{W,Z,\Theta} - \bar{Q}^1_\Sigma)^{-1} = \mathcal{G}_{\bar{Q}^1_{W,Z,\Theta}+\bar{Q}^1_\Sigma}(\bar{Z}^1).$$

Further by the subordination formula (12),

$$\bar{G}^1 = \mathcal{G}_{\bar{Q}^1_\Sigma}(\bar{Z}^1 - \mathcal{R}_{\bar{Q}^1_{W,Z,\Theta}}(\bar{G}^1)) = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^1 - \mathcal{R}_{\bar{Q}^1_{W,Z,\Theta}}(\bar{G}^1) - \bar{Q}^1_\Sigma)^{-1}. \tag{22}$$

Since $\bar{Q}^1_{W,Z,\Theta}$ consists of i.i.d. Gaussian blocks, by (13), its limiting R-transform has a form

$$\mathcal{R}_{\bar{Q}^1_{W,Z,\Theta}}(\bar{G}^1) = \begin{bmatrix} \cdot & (R^1)^\top \\ R^1 & \cdot \end{bmatrix},$$

where the non-zero blocks of R^1 are

$$R^1_{1,1} = -\frac{\rho_s\omega_s}{\gamma}G^1_{2,2} - \frac{\sqrt{\rho_s}}{\gamma}G^1_{3,6}, \quad R^1_{1,7} = -\frac{\sqrt{\rho_s}}{\gamma}G^1_{3,12}, \quad R^1_{2,2} = -\frac{\psi\rho_s\omega_s}{\gamma\phi}G^1_{1,1} + \sqrt{\rho_s}\psi G^1_{5,4},$$
$$R^1_{2,14} = \sqrt{\rho_s}\psi G^1_{5,13}, \quad R^1_{4,5} = \sqrt{\rho_s}G^1_{2,2}, \quad R^1_{6,3} = -\frac{\sqrt{\rho_s}}{\gamma\phi}G^1_{1,1}, \quad R^1_{6,9} = -\frac{\sqrt{\rho_s}}{\gamma\phi}G^1_{1,7},$$
$$R^1_{7,1} = -\frac{\sqrt{\rho_s}}{\gamma}G^1_{9,6} = 0, \quad R^1_{7,7} = -\frac{\rho_s\omega_s}{\gamma}G^1_{8,8} - \frac{\sqrt{\rho_s}}{\gamma}G^1_{9,12}, \quad R^1_{8,8} = -\frac{\psi\rho_s\omega_s}{\gamma\phi}G^1_{7,7} + \sqrt{\rho_s}\psi G^1_{11,10},$$
$$R^1_{10,11} = \sqrt{\rho_s}G^1_{8,8}, \quad R^1_{12,3} = -\frac{\sqrt{\rho_s}}{\gamma\phi}G^1_{7,1} = 0, \quad R^1_{12,9} = -\frac{\sqrt{\rho_s}}{\gamma\phi}G^1_{7,7}, \quad R^1_{13,5} = \sqrt{\rho_s}G^1_{14,2} = 0.$$

We used the fact that $G^1_{9,6} = G^1_{7,1} = G^1_{14,2} = 0$, which we obtain from block matrix inversion of Q^1.

Computing the block-matrix inverse of Q^1 and from equations (19), (21), we see

$$G^1_{1,1} = G^1_{7,7} = \gamma\mathbb{E}\overline{\mathrm{tr}}(K^{-1}) = G^0_{1,1} = \gamma\tau, \quad G^1_{2,2} = G^1_{8,8} = \gamma\mathbb{E}\overline{\mathrm{tr}}(\hat{K}^{-1}) = G^0_{2,2} = \gamma\bar{\tau},$$
$$G^1_{3,6} = G^1_{9,12} = \frac{\gamma\sqrt{\rho_s}\,\mathbb{E}\overline{\mathrm{tr}}[\Sigma_s W^\top \hat{K}^{-1} W]}{N} = G^0_{3,6} = \gamma\sqrt{\rho_s}\bar{\tau}\mathcal{I}^s_{1,1},$$
$$G^1_{5,4} = G^1_{11,10} = -\frac{\sqrt{\rho_s}\,\mathbb{E}\overline{\mathrm{tr}}[\Sigma_s Z K^{-1} Z^\top]}{d} = G^0_{5,4} = -\frac{\sqrt{\rho_s}\tau\mathcal{I}^s_{1,1}}{\phi}.$$

Plugging these into (22), we obtain self-consistent equations. For example,

$$\begin{aligned} G^1_{2,14} &= \mathbb{E}\overline{\mathrm{tr}}[(I_{2n+9d+3N} - R^1 - Q^1_\Sigma)^{-1}]_{2,14} \\ &= -\frac{\gamma\sqrt{\rho_s}\psi\phi G^1_{5,13}}{\gamma\phi(-1+\sqrt{\rho_s}\psi G^1_{5,4}) - \psi\rho_s\omega_s G^1_{1,1}} = \gamma\sqrt{\rho_s}\bar{\tau}\psi G^1_{5,13}. \end{aligned}$$

Similarly,

$$G^1_{5,13} = \mathbb{E}_\mu\left[\frac{\lambda^s\lambda^j\gamma\sqrt{\rho_s}\phi G^1_{1,7}G^1_{2,2} + (\lambda^s)^2\lambda^j\sqrt{\rho_s}(G^1_{1,1})^2G^1_{2,2}}{(\gamma\phi + \lambda^s\rho_s G^1_{1,1}G^1_{2,2})^2}\right] = \sqrt{\rho_s}\bar{\tau}\mathcal{I}^j_{2,2}G^1_{1,7} + \frac{\gamma\sqrt{\rho_s}\tau^2\bar{\tau}}{\phi}\mathcal{I}^j_{3,2},$$
$$G^1_{1,7} = -\frac{\gamma\sqrt{\rho_s}G^1_{3,12}}{(\gamma + \sqrt{\rho_s}G^1_{3,6} + \rho_s\omega_s G^1_{2,2})^2} = -\gamma\sqrt{\rho_s}\tau^2 G^1_{3,12},$$
$$G^1_{3,12} = -\mathbb{E}_\mu\left[\frac{\lambda^s\gamma^2\phi^2 + (\lambda^s)^2\gamma\rho_s^2\phi G^1_{1,7}(G^1_{2,2})^2}{\sqrt{\rho_s}(\gamma\phi + \lambda^s\rho_s G^1_{1,1}G^1_{2,2})^2}\right] = -\frac{\phi}{\sqrt{\rho_s}}\mathcal{I}^s_{1,2} - \gamma\rho_s^{\frac{3}{2}}\bar{\tau}^2\mathcal{I}^s_{2,2}G^1_{1,7}.$$

Eliminating $G^1_{3,12}$ and using $\kappa = \gamma\rho_s\tau\bar{\tau}$,

$$G^1_{1,7} = \gamma\tau^2\phi\mathcal{I}^s_{1,2} + \kappa^2\mathcal{I}^s_{2,2}G^1_{1,7} \Rightarrow G^1_{1,7} = \frac{\gamma\tau^2\phi\mathcal{I}^s_{1,2}}{1-\kappa^2\mathcal{I}^s_{2,2}}.$$

Therefore,

$$\begin{aligned} G^1_{2,14} &= \gamma\sqrt{\rho_s}\bar{\tau}\psi G^1_{5,13} = \gamma\rho_s\bar{\tau}^2\psi\mathcal{I}^j_{2,2}G^1_{1,7} + \frac{\gamma^2\rho_s\tau^2\bar{\tau}^2\psi}{\phi}\mathcal{I}^j_{3,2} \\ &= \frac{\gamma^2\rho_s\tau^2\bar{\tau}^2\psi\phi\mathcal{I}^s_{1,2}\mathcal{I}^j_{2,2}}{1-\kappa^2\mathcal{I}^s_{2,2}} + \frac{\gamma^2\rho_s\tau^2\bar{\tau}^2\psi}{\phi}\mathcal{I}^j_{3,2} = \kappa^2\left(\frac{\psi\phi\mathcal{I}^s_{1,2}\mathcal{I}^j_{2,2}}{\rho_s(1-\kappa^2\mathcal{I}^s_{2,2})} + \frac{\psi\mathcal{I}^j_{3,2}}{\rho_s\phi}\right). \end{aligned}$$

Finally,

$$D_{21} = \frac{2\rho_j}{\psi}G^1_{2,14} = \frac{2\rho_j\kappa^2}{\rho_s}\left(\frac{\phi\mathcal{I}^s_{1,2}\mathcal{I}^j_{2,2}}{1-\kappa^2\mathcal{I}^s_{2,2}} + \frac{\mathcal{I}^j_{3,2}}{\phi}\right). \tag{23}$$

B.1.4. COMPUTATION OF D_{22}

Let

$$Q^2 = \begin{bmatrix}
I_n & \frac{\sqrt{\rho_s\omega_s}\Theta_2^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{\sqrt{\rho_s\omega_s}\Theta_2}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_s}W_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & I_d & -\Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -\frac{W_2^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_n & \frac{\sqrt{\rho_s\omega_s}\Theta_1^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_1}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_s}W_1}{\sqrt{N}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_s^{\frac{1}{2}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{W_1^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot \\
\cdot & \cdot & \cdot & \Sigma_j & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_s^{\frac{1}{2}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d
\end{bmatrix}$$

and $G^2 = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((Q^2)^{-1})$. Then,

$$G_{7,1}^2 = \frac{\gamma\sqrt{\rho_s}\phi}{dN^2}\mathbb{E}\,\mathrm{tr}[K_1^{-1}F_1^\top W_1\Sigma_j W_2^\top F_2 K_2^{-1}] = \frac{\gamma\sqrt{\rho_s}\phi}{2\sigma_\varepsilon^2\rho_j}D_{22}.$$

We augment Q^2 to the symmetric matrix $\bar{Q}^2$ as

$$\bar{Q}^2 = \begin{bmatrix} \cdot & (Q^2)^\top \\ Q^2 & \cdot \end{bmatrix}$$

and write

$$\begin{aligned}
\bar{Q}^2 &= \bar{Z}^2 - \bar{Q}^2_{W,Z,\Theta} - \bar{Q}^2_\Sigma \\
&= \begin{bmatrix} \cdot & I_{2n+8d+2N} \\ I_{2n+8d+2N} & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & (Q^2_{W,Z,\Theta})^\top \\ Q^2_{W,Z,\Theta} & \cdot \end{bmatrix} - \begin{bmatrix} \cdot & (Q^2_\Sigma)^\top \\ Q^2_\Sigma & \cdot \end{bmatrix},
\end{aligned}$$

where

$$Q^2_{W,Z,\Theta} = \begin{bmatrix}
\cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_2^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{\sqrt{\rho_s\omega_s}\Theta_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \frac{W_2^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_1^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s\omega_s}\Theta_1}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W_1}{\sqrt{N}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W_1^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{Z}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot
\end{bmatrix},$$

and

$$Q_\Sigma^2 = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -\Sigma_j & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

Defining $\bar{G}^2$ below,

$$\begin{aligned} \bar{G}^2 &= \begin{bmatrix} \cdot & G^2 \\ (G^2)^\top & \cdot \end{bmatrix} = \begin{bmatrix} \cdot & (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((Q^2)^{-1}) \\ (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})(((Q^2)^\top)^{-1}) & \cdot \end{bmatrix} \\ &= (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}}) \begin{bmatrix} \cdot & (Q^2)^{-1} \\ ((Q^2)^\top)^{-1} & \cdot \end{bmatrix} = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((\bar{Q}^2)^{-1}). \end{aligned}$$

It can be viewed as the operator-valued Cauchy transform of $\bar{Q}^2_{W,Z,\Theta} + \bar{Q}^2_\Sigma$ (in the space we consider in Remark A.7), i.e.,

$$\bar{G}^2 = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^2 - \bar{Q}^2_{W,Z,\Theta} - \bar{Q}^2_\Sigma)^{-1} = \mathcal{G}_{\bar{Q}^2_{W,Z,\Theta} + \bar{Q}^2_\Sigma}(\bar{Z}^2).$$

Further by the subordination formula (12),

$$\bar{G}^2 = \mathcal{G}_{\bar{Q}^2_\Sigma}(\bar{Z}^2 - \mathcal{R}_{\bar{Q}^2_{W,Z,\Theta}}(\bar{G}^2)) = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^2 - \mathcal{R}_{\bar{Q}^2_{W,Z,\Theta}}(\bar{G}^2) - \bar{Q}^2_\Sigma)^{-1}. \tag{24}$$

Since $\bar{Q}^2_{W,Z,\Theta}$ consists of i.i.d. Gaussian blocks, by (13), its limiting R-transform has a form

$$\mathcal{R}_{\bar{Q}^2_{W,Z,\Theta}}(\bar{G}^2) = \begin{bmatrix} \cdot & (R^2)^\top \\ R^2 & \cdot \end{bmatrix},$$

where the non-zero blocks of R^2 are

$$\begin{aligned} &R^2_{1,1} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma} G^2_{2,2} - \frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^2_{3,6}, \quad R^2_{1,7} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^2_{3,12} = 0, \quad R^2_{2,2} = -\frac{\psi\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma\phi} G^2_{1,1} + \sqrt{\rho_{\mathrm{s}}}\psi G^2_{5,4}, \\ &R^2_{4,5} = \sqrt{\rho_{\mathrm{s}}} G^2_{2,2}, \quad R^2_{6,3} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^2_{1,1}, \quad R^2_{6,9} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^2_{1,7} = 0, \quad R^2_{7,1} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^2_{9,6}, \\ &R^2_{7,7} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma} G^2_{8,8} - \frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^2_{9,12}, \quad R^2_{8,8} = -\frac{\psi\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma\phi} G^2_{7,7} + \sqrt{\rho_{\mathrm{s}}}\psi G^2_{11,10}, \quad R^2_{10,11} = \sqrt{\rho_{\mathrm{s}}} G^2_{8,8}, \\ &R^2_{12,3} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^2_{7,1}, \quad R^2_{12,9} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^2_{7,7}. \end{aligned}$$

We used the fact that $G^2_{3,12} = G^2_{1,7} = 0$, which we obtain from block matrix inversion of Q^2. From block matrix inversion of Q^2 and equations (19), (21), we have

$$\begin{aligned} &G^2_{1,1} = G^2_{7,7} = \gamma\mathbb{E}\overline{\mathrm{tr}}(K^{-1}) = G^0_{1,1} = \gamma\tau, \quad G^2_{2,2} = G^2_{8,8} = \gamma\mathbb{E}\overline{\mathrm{tr}}(\hat{K}^{-1}) = G^0_{2,2} = \gamma\bar{\tau}, \\ &G^2_{3,6} = G^2_{9,12} = \frac{\gamma\sqrt{\rho_{\mathrm{s}}}\,\mathbb{E}\overline{\mathrm{tr}}[\Sigma_{\mathrm{s}} W^\top \hat{K}^{-1} W]}{N} = G^0_{3,6} = \gamma\sqrt{\rho_{\mathrm{s}}}\bar{\tau}\mathcal{I}^{\mathrm{s}}_{1,1}, \\ &G^2_{5,4} = G^2_{11,10} = -\frac{\sqrt{\rho_{\mathrm{s}}}\,\mathbb{E}\overline{\mathrm{tr}}[\Sigma_{\mathrm{s}} Z K^{-1} Z^\top]}{d} = G^0_{5,4} = -\frac{\sqrt{\rho_{\mathrm{s}}}\tau\mathcal{I}^{\mathrm{s}}_{1,1}}{\phi}. \end{aligned}$$

Plugging these into (24), we have the following self-consistent equations

$$G_{7,1}^2 = -\frac{\gamma\sqrt{\rho_{\rm s}}G_{9,6}^2}{(\gamma+\sqrt{\rho_{\rm s}}G_{3,6}^2+\rho_{\rm s}\omega_{\rm s}G_{2,2}^2)^2} = -\gamma\sqrt{\rho_{\rm s}}\tau^2 G_{9,6}^2,$$
$$G_{9,6}^2 = -\mathbb{E}_\mu\left[\frac{(\lambda^{\rm s})^2\gamma\rho_{\rm s}^{\frac{3}{2}}\phi(G_{2,2}^2)^2G_{7,1}^2+\lambda^{\rm s}\lambda^j\gamma^2\rho_{\rm s}\phi^2(G_{2,2}^2)^2}{(\gamma\phi+\lambda^{\rm s}\rho_{\rm s}G_{1,1}^2G_{2,2}^2)^2}\right] = -\gamma\rho_{\rm s}^{\frac{3}{2}}\bar{\tau}^2\mathcal{I}_{2,2}^{\rm s}G_{7,1}^2-\gamma^2\rho_{\rm s}\bar{\tau}^2\phi\mathcal{I}_{2,2}^j.$$

Solving for $G_{7,1}^2$,

$$G_{7,1}^2 = \frac{\kappa^2\gamma\phi\mathcal{I}_{2,2}^j}{\sqrt{\rho_{\rm s}}(1-\kappa^2\mathcal{I}_{2,2}^{\rm s})}.$$

Therefore,

$$D_{22} = \frac{2\sigma_\varepsilon^2\rho_j}{\gamma\sqrt{\rho_{\rm s}}\phi}G_{7,1}^2 = \frac{2\rho_j\kappa^2\sigma_\varepsilon^2\mathcal{I}_{2,2}^j}{\rho_{\rm s}(1-\kappa^2\mathcal{I}_{2,2}^{\rm s})}. \tag{25}$$

B.1.5. Computation of $\mathrm{Dis}_{\rm SS}^j(\phi,\psi,\gamma)$

Combining equations (16), (17), (18), (23), (25), we get

$$\mathrm{Dis}_{\rm SS}^j(\phi,\psi,\gamma) = \mathrm{Dis}_{\rm I}^j(\phi,\psi,\gamma) - \frac{2\rho_j\kappa^2(\sigma_\varepsilon^2+\phi\mathcal{I}_{1,2}^{\rm s})\mathcal{I}_{2,2}^j}{\rho_{\rm s}(1-\kappa^2\mathcal{I}_{2,2}^{\rm s})}.$$

B.1.6. Decomposition of $\mathrm{Dis}_{\rm SW}^j(\phi,\psi,\gamma)$

Writing $F_i = \sigma(WX_i/\sqrt{d})$, $f=\sigma(Wx/\sqrt{d})$, $K_i = \frac{1}{N}F_i^\top F_i+\gamma I_n$ for $i\in\{1,2\}$, we can write SW disagreement as

$$\begin{aligned}\mathrm{Dis}_{\rm SW}^j(\phi,\psi,\gamma) &= \frac{1}{N^2}\mathbb{E}[(Y_1^\top K_1^{-1}F_1^\top f - Y_2^\top K_2^{-1}F_2^\top f)^2]\\ &= \frac{2}{N^2}\mathbb{E}[f^\top F_1K_1^{-1}Y_1Y_1^\top K_1^{-1}F_1^\top f] - \frac{2}{N^2}\mathbb{E}[f^\top F_2K_2^{-1}Y_2Y_1^\top K_1^{-1}F_1^\top f]\\ &= D_1 - D_3.\end{aligned} \tag{26}$$

The term D_1 is given in (17). Plugging in $Y_i = X_i^\top\beta/\sqrt{d}+\boldsymbol{\varepsilon}_i$, where $\boldsymbol{\varepsilon}_i=(\varepsilon_{i1},\dots,\varepsilon_{in})^\top\in\mathbb{R}^n$, the term D_3 becomes

$$\begin{aligned}D_3 =& \frac{2}{dN^2}\mathbb{E}_{W,X_i}\,\mathrm{tr}[F_2K_2^{-1}X_2^\top\mathbb{E}_\beta[\beta\beta^\top]X_1K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[ff^\top]]\\ &+\frac{4}{\sqrt{d}N^2}\mathbb{E}_{W,X_i}[F_2K_2^{-1}X_2^\top\mathbb{E}_{\beta,\boldsymbol{\varepsilon}_1}[\beta\boldsymbol{\varepsilon}_1^\top]K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[ff^\top]]\\ &+\frac{2}{N^2}\mathbb{E}_{W,X_i}\,\mathrm{tr}[F_2K_2^{-1}\mathbb{E}_{\boldsymbol{\varepsilon}_i}[\boldsymbol{\varepsilon}_2\boldsymbol{\varepsilon}_1^\top]K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[ff^\top]]\\ =&\frac{2}{dN^2}\mathbb{E}_{W,X_i}\,\mathrm{tr}[F_2K_2^{-1}X_2^\top X_1K_1^{-1}F_1^\top\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[ff^\top]].\end{aligned}$$

From the Gaussian equivalence (14), we have

$$\mathbb{E}_{x\sim\mathcal{D}_j,\theta}[ff^\top] = \frac{\rho_j}{d}W\Sigma_jW^\top+\rho_j\omega_jI_N.$$

Therefore,

$$\begin{aligned}D_3 &= \frac{2\rho_j}{d^2N^2}\mathbb{E}_{W,X_i}\,\mathrm{tr}[W\Sigma_jW^\top F_2K_2^{-1}X_2^\top X_1K_1^{-1}F_1^\top] + \frac{2\rho_j\omega_j}{dN^2}\mathbb{E}_{W,X_i}\,\mathrm{tr}\left[F_2K_2^{-1}X_2^\top X_1K_1^{-1}F_1^\top\right]\\ &= D_{31}+D_{32}.\end{aligned} \tag{27}$$

We can write $X_i = \Sigma_{\mathrm{s}}^{\frac{1}{2}} Z_i$ for $Z_i \in \mathbb{R}^{d\times n}$ with i.i.d. standard Gaussian entries. Thus,

$$D_{31} = \frac{2\rho_j}{d^2N^2}\mathbb{E}_{W,Z_i}\operatorname{tr}[W\Sigma_j W^\top F_2 K_2^{-1} Z_2^\top \Sigma_{\mathrm{s}} Z_1 K_1^{-1} F_1^\top],$$
$$D_{32} = \frac{2\rho_j\omega_j}{dN^2}\mathbb{E}_{W,Z_i}\operatorname{tr}\left[F_2 K_2^{-1} Z_2^\top \Sigma_{\mathrm{s}} Z_1 K_1^{-1} F_1^\top\right].$$

Now, we use the linear pencil method to compute D_{31} and D_{32}.

B.1.7. Computation of D_{31}

Let

$$Q^3 = \begin{bmatrix}
I_n & \frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_2^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\mathrm{s}}}Z_2^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_2}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\mathrm{s}}}W}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -\frac{W^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\Sigma_{\mathrm{s}}^{\frac{1}{2}}}{\sqrt{\rho_{\mathrm{s}}}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{Z_2}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_n & \frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_1^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\mathrm{s}}}Z_1^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_1}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\mathrm{s}}}W}{\sqrt{N}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{W^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\mathrm{s}}^{\frac{1}{2}}\frac{\Sigma_j}{\sqrt{\rho_{\mathrm{s}}}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{Z_1}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\frac{W^\top}{\sqrt{N}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_N
\end{bmatrix}$$

and $G^3 = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((Q^3)^{-1})$. Then,

$$G^3_{2,14} = \frac{\psi}{d^2N^2}\mathbb{E}_{W,Z_i}\operatorname{tr}[W\Sigma_j W^\top F_2 K_2^{-1} Z_2^\top \Sigma_{\mathrm{s}} Z_1 K_1^{-1} F_1^\top] = \frac{\psi}{2\rho_j} D_{31}.$$

We augment Q^3 to the symmetric matrix $\bar{Q}^3$ as

$$\bar{Q}^3 = \begin{bmatrix} 0 & (Q^3)^\top \\ Q^3 & 0 \end{bmatrix}$$

and write

$$\begin{aligned}\bar{Q}^3 &= \bar{Z}^3 - \bar{Q}^3_{W,Z,\Theta} - \bar{Q}^3_{\Sigma} \\ &= \begin{bmatrix} 0 & I_{2n+9d+3N} \\ I_{2n+9d+3N} & 0 \end{bmatrix} - \begin{bmatrix} 0 & (Q^3_{W,Z,\Theta})^\top \\ Q^3_{W,Z,\Theta} & 0 \end{bmatrix} - \begin{bmatrix} 0 & (Q^3_{\Sigma})^\top \\ Q^3_{\Sigma} & 0 \end{bmatrix},\end{aligned}$$

where

$$Q^3_{W,Z,\Theta} = \begin{bmatrix}
\cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_2^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z_2^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{\sqrt{\rho_s\omega_s}\Theta_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \frac{W^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{Z_2}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_s\omega_s}\Theta_1^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_s}Z_1^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s\omega_s}\Theta_1}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_s}W}{\sqrt{N}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{Z_1}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W^\top}{\sqrt{N}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{bmatrix}$$

and

$$Q^3_\Sigma = \begin{bmatrix}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\Sigma_s^{\frac{1}{2}}}{\sqrt{\rho_s}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_s^{\frac{1}{2}} & -\frac{\Sigma_j}{\sqrt{\rho_s}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{bmatrix}.$$

Defining $\bar{G}^3$ below,

$$\begin{aligned}
\bar{G}^3 &= \begin{bmatrix} 0 & G^3 \\ (G^3)^\top & 0 \end{bmatrix} = \begin{bmatrix} 0 & (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((Q^3)^{-1}) \\ (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(((Q^3)^\top)^{-1}) & 0 \end{bmatrix} \\
&= (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}}) \begin{bmatrix} 0 & (Q^3)^{-1} \\ ((Q^3)^\top)^{-1} & 0 \end{bmatrix} = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})((\bar{Q}^3)^{-1}).
\end{aligned}$$

It can be viewed as the operator-valued Cauchy transform of $\bar{Q}^3_{W,Z,\Theta} + \bar{Q}^3_\Sigma$ (in the space we consider in Remark A.7), i.e.,

$$\bar{G}^3 = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^3 - \bar{Q}^3_{W,Z,\Theta} - \bar{Q}^3_\Sigma)^{-1} = \mathcal{G}_{\bar{Q}^3_{W,Z,\Theta}+\bar{Q}^3_\Sigma}(\bar{Z}^3).$$

Further by the subordination formula (12),

$$\bar{G}^3 = \mathcal{G}_{\bar{Q}^3_\Sigma}(\bar{Z} - \mathcal{R}_{\bar{Q}^3_{W,Z,\Theta}}(\bar{G}^3)) = (\mathrm{id}\otimes\mathbb{E}\overline{\mathrm{tr}})(\bar{Z}^3 - \mathcal{R}_{\bar{Q}^3_{W,Z,\Theta}}(\bar{G}^3) - \bar{Q}^3_\Sigma)^{-1}. \tag{28}$$

Since $\bar{Q}^3_{W,Z,\Theta}$ consists of i.i.d. Gaussian blocks, by (13), its limiting R-transform has a form

$$\mathcal{R}_{\bar{Q}^3_{W,Z,\Theta}}(\bar{G}^3) = \begin{bmatrix} 0 & (R^3)^\top \\ R^3 & 0 \end{bmatrix},$$

where the non-zero blocks of R^3 are

$$R^3_{1,1} = -\frac{\rho_{\rm s}\omega_{\rm s}}{\gamma}G^3_{2,2} - \frac{\sqrt{\rho_{\rm s}}}{\gamma}G^3_{3,6}, \quad R^3_{2,2} = -\frac{\psi\rho_{\rm s}\omega_{\rm s}}{\gamma\phi}G^3_{1,1} + \sqrt{\rho_{\rm s}}\psi G^3_{5,4}, \quad R^3_{2,8} = \sqrt{\rho_{\rm s}}\psi G^3_{5,10}, \quad R^3_{2,14} = \sqrt{\rho_{\rm s}}\psi G^3_{5,13},$$
$$R^3_{4,5} = \sqrt{\rho_{\rm s}}G^3_{2,2}, \quad R^3_{4,11} = \sqrt{\rho_{\rm s}}G^3_{2,8}, \quad R^3_{6,3} = -\frac{\sqrt{\rho_{\rm s}}}{\gamma\phi}G^3_{1,1}, \quad R^3_{7,7} = -\frac{\rho_{\rm s}\omega_{\rm s}}{\gamma}G^3_{8,8} - \frac{\sqrt{\rho_{\rm s}}}{\gamma}G^3_{9,12},$$
$$R^3_{8,2} = \sqrt{\rho_{\rm s}}\psi G^3_{11,4} = 0, \quad R^3_{8,8} = -\frac{\psi\rho_{\rm s}\omega_{\rm s}}{\gamma\phi}G^3_{7,7} + \sqrt{\rho_{\rm s}}\psi G^3_{11,10}, \quad R^3_{8,14} = \sqrt{\rho_{\rm s}}\psi G^3_{11,13}, \quad R^3_{10,5} = \sqrt{\rho_{\rm s}}G^3_{8,2} = 0,$$
$$R^3_{10,11} = \sqrt{\rho_{\rm s}}G^3_{8,8}, \quad R^3_{12,9} = -\frac{\sqrt{\rho_{\rm s}}}{\gamma\phi}G^3_{7,7}, \quad R^3_{13,5} = \sqrt{\rho_{\rm s}}G^3_{14,2} = 0, \quad R^3_{13,11} = \sqrt{\rho_{\rm s}}G^3_{14,8} = 0.$$

We used the fact that $G^3_{11,4} = G^3_{8,2} = G^3_{14,2} = G^3_{14,8} = 0$, which we obtain from block matrix inversion of Q^3.

Further from block matrix inversion of Q^3 and equations (19), (21), we have

$$G^3_{1,1} = G^3_{7,7} = \gamma\mathbb{E}\overline{\rm tr}(K^{-1}) = G^0_{1,1} = \gamma\tau, \quad G^3_{2,2} = G^3_{8,8} = \gamma\mathbb{E}\overline{\rm tr}(\hat{K}^{-1}) = G^0_{2,2} = \gamma\bar{\tau},$$
$$G^3_{3,6} = G^3_{9,12} = \frac{\gamma\sqrt{\rho_{\rm s}}\,\mathbb{E}\overline{\rm tr}[\Sigma_{\rm s}W^\top\hat{K}^{-1}W]}{N} = G^0_{3,6} = \gamma\sqrt{\rho_{\rm s}}\bar{\tau}\mathcal{I}^{\rm s}_{1,1},$$
$$G^3_{5,4} = G^3_{11,10} = -\frac{\sqrt{\rho_{\rm s}}\,\mathbb{E}\overline{\rm tr}[\Sigma_{\rm s}ZK^{-1}Z^\top]}{d} = G^0_{5,4} = -\frac{\sqrt{\rho_{\rm s}}\tau\mathcal{I}^{\rm s}_{1,1}}{\phi}.$$

Plugging these into (28), we have the following self-consistent equations

$$G^3_{2,14} = \gamma^2\rho_{\rm s}\bar{\tau}^2\psi^2G^3_{5,10}G^3_{11,13} + \gamma\sqrt{\rho_{\rm s}}\bar{\tau}\psi G^3_{5,13}, \quad G^3_{5,10} = -\frac{\sqrt{\rho_{\rm s}}\tau^2}{\phi}\mathcal{I}^{\rm s}_{2,2} + \frac{\rho_{\rm s}^{\frac{3}{2}}\tau^2}{\phi}\mathcal{I}^{\rm s}_{2,2}G^3_{2,8},$$
$$G^3_{2,8} = \gamma^2\sqrt{\rho_{\rm s}}\bar{\tau}^2\psi G^3_{5,10}, \quad G^3_{5,13} = \sqrt{\rho_{\rm s}}\tau\mathcal{I}^j_{2,2}G^3_{2,8} + \frac{\gamma\sqrt{\rho_{\rm s}}\tau^2\bar{\tau}}{\phi}\mathcal{I}^j_{3,2}, \quad G^3_{11,13} = -\frac{\mathcal{I}^j_{1,1}}{\sqrt{\rho_{\rm s}}}.$$

Solving for $G^3_{5,10}$ gives

$$G^3_{5,10} = -\frac{\sqrt{\rho_{\rm s}}\tau^2\mathcal{I}^{\rm s}_{2,2}}{\phi - \psi\kappa^2\mathcal{I}^{\rm s}_{2,2}}.$$

Plugging in $G^3_{5,10}, G^3_{11,13}, G^3_{5,13}$ to find $G^3_{2,14}$, we get

$$D_{31} = \frac{2\rho_j}{\psi}G^3_{2,14} = \frac{2\rho_j\psi\phi\kappa^2\mathcal{I}^{\rm s}_{2,2}\mathcal{I}^j_{1,2}}{\rho_{\rm s}(\phi - \psi\kappa^2\mathcal{I}^{\rm s}_{2,2})} + \frac{2\rho_j\kappa^2}{\rho_{\rm s}\phi}\mathcal{I}^j_{3,2}. \tag{29}$$

B.1.8. COMPUTATION OF D_{32}

Let

$$Q^4 = \begin{bmatrix}
I_n & \frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_2^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\rm s}}Z_2^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_2}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}}W}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & -\frac{W^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & \cdot \\
\cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
-\frac{Z_2}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_n & \frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_1^\top}{\gamma\sqrt{N}} & \frac{\sqrt{\rho_{\rm s}}Z_1^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}\omega_{\rm s}}\Theta_1}{\sqrt{N}} & I_N & \cdot & \cdot & -\frac{\sqrt{\rho_{\rm s}}W}{\sqrt{N}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac{1}{2}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{W^\top}{\sqrt{N}} & \cdot & I_d & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & -\Sigma_{\rm s}^{\frac{1}{2}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{Z_1}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & I_d
\end{bmatrix}$$

and $G^4 = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((Q^4)^{-1})$. Then,

$$G^4_{2,8} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{dN^2}\mathbb{E}_{W,Z_i}\,\mathrm{tr}\left[F_2 K_2^{-1} Z_2^\top \Sigma_{\mathrm{s}} Z_1 K_1^{-1} F_1^\top\right] = -\frac{\sqrt{\rho_{\mathrm{s}}}}{2\rho_j\omega_j} D_{32}.$$

We augment Q^4 to the symmetric matrix $\bar{Q}^4$ as

$$\bar{Q}^4 = \begin{bmatrix} 0 & (Q^4)^\top \\ Q^4 & 0 \end{bmatrix}$$

and write

$$\begin{aligned}\bar{Q}^4 &= \bar{Z}^4 - \bar{Q}^4_{W,Z,\Theta} - \bar{Q}^4_{\Sigma} \\ &= \begin{bmatrix} 0 & I_{2n+8d+2N} \\ I_{2n+8d+2N} & 0 \end{bmatrix} - \begin{bmatrix} 0 & (Q^4_{W,Z,\Theta})^\top \\ Q^4_{W,Z,\Theta} & 0 \end{bmatrix} - \begin{bmatrix} 0 & (Q^4_{\Sigma})^\top \\ Q^4_{\Sigma} & 0 \end{bmatrix},\end{aligned}$$

where

$$Q^4_{W,Z,\Theta} = \begin{bmatrix}
\cdot & -\frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_2^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_{\mathrm{s}}}Z_2^\top}{\gamma\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_2}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_{\mathrm{s}}}W}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \frac{W^\top}{\sqrt{N}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\frac{Z_2}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & -\frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_1^\top}{\gamma\sqrt{N}} & -\frac{\sqrt{\rho_{\mathrm{s}}}Z_1^\top}{\gamma\sqrt{d}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}\Theta_1}{\sqrt{N}} & \cdot & \cdot & \cdot & \frac{\sqrt{\rho_{\mathrm{s}}}W}{\sqrt{N}} & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \frac{W^\top}{\sqrt{N}} & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \frac{Z_1}{\sqrt{d}} & \cdot & \cdot & \cdot & \cdot & \cdot
\end{bmatrix}$$

and

$$Q^4_{\Sigma} = \begin{bmatrix}
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & I_d & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \Sigma_{\mathrm{s}}^{\frac{1}{2}} \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot
\end{bmatrix}.$$

Defining $\bar{G}^4$ below,

$$\begin{aligned}\bar{G}^4 &= \begin{bmatrix} 0 & G^4 \\ (G^4)^\top & 0 \end{bmatrix} = \begin{bmatrix} 0 & (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((Q^4)^{-1}) \\ (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})(((Q^4)^\top)^{-1}) & 0 \end{bmatrix} \\ &= (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}}) \begin{bmatrix} 0 & (Q^4)^{-1} \\ ((Q^4)^\top)^{-1} & 0 \end{bmatrix} = (\mathrm{id} \otimes \mathbb{E}\overline{\mathrm{tr}})((\bar{Q}^4)^{-1}).\end{aligned}$$

It can be viewed as the operator-valued Cauchy transform of $\bar{Q}^4_{W,Z,\Theta} + \bar{Q}^4_\Sigma$ (in the space we consider in Remark A.7), i.e.,

$$\bar{G}^4 = (\mathrm{id} \otimes \overline{\mathbb{E}\mathrm{tr}})(\bar{Z}^4 - \bar{Q}^4_{W,Z,\Theta} - \bar{Q}^4_\Sigma)^{-1} = \mathcal{G}_{\bar{Q}^4_{W,Z,\Theta}+\bar{Q}^4_\Sigma}(\bar{Z}^4).$$

Further by the subordination formula (12),

$$\bar{G}^4 = \mathcal{G}_{\bar{Q}^4_\Sigma}(\bar{Z} - \mathcal{R}_{\bar{Q}^4_{W,Z,\Theta}}(\bar{G}^4)) = (\mathrm{id} \otimes \overline{\mathbb{E}\mathrm{tr}})(\bar{Z}^4 - \mathcal{R}_{\bar{Q}^4_{W,Z,\Theta}}(\bar{G}^4) - \bar{Q}^4_\Sigma)^{-1}. \tag{30}$$

Since $\bar{Q}^4_{W,Z,\Theta}$ consists of i.i.d. Gaussian blocks, by (13), its limiting R-transform has a form

$$\mathcal{R}_{\bar{Q}^4_{W,Z,\Theta}}(\bar{G}^4) = \begin{bmatrix} 0 & (R^4)^\top \\ R^4 & 0 \end{bmatrix},$$

where the non-zero blocks of R^4 are

$$R^4_{1,1} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma} G^4_{2,2} - \frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^4_{3,6}, \quad R^4_{2,2} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}\psi}{\gamma\phi} G^4_{1,1} + \sqrt{\rho_{\mathrm{s}}}\psi G^4_{5,4}, \quad R^4_{2,8} = \sqrt{\rho_{\mathrm{s}}}\psi G^4_{5,10}, \quad R^4_{4,5} = \sqrt{\rho_{\mathrm{s}}} G^4_{2,2}$$
$$R^4_{4,11} = \sqrt{\rho_{\mathrm{s}}} G^4_{2,8}, \quad R^4_{6,3} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^4_{1,1}, \quad R^4_{7,7} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}}{\gamma} G^4_{8,8} - \frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma} G^4_{9,12}, \quad R^4_{8,2} = \sqrt{\rho_{\mathrm{s}}}\psi G^4_{11,4} = 0,$$
$$R^4_{8,8} = -\frac{\rho_{\mathrm{s}}\omega_{\mathrm{s}}\psi}{\gamma\phi} G^4_{7,7} + \sqrt{\rho_{\mathrm{s}}}\psi G^4_{11,10}, \quad R^4_{10,5} = \sqrt{\rho_{\mathrm{s}}} G^4_{8,2} = 0, \quad R^4_{10,11} = \sqrt{\rho_{\mathrm{s}}} G^4_{8,8}, \quad R^4_{12,9} = -\frac{\sqrt{\rho_{\mathrm{s}}}}{\gamma\phi} G^4_{7,7}.$$

We used the fact that $G^4_{11,4} = G^4_{8,2} = 0$, which we obtain from block matrix inversion of Q^4.

Further from block matrix inversion of Q^4 and equations (19), (21), we have

$$G^4_{1,1} = G^4_{7,7} = \gamma\overline{\mathbb{E}\mathrm{tr}}(K^{-1}) = G^0_{1,1} = \gamma\tau, \quad G^4_{2,2} = G^4_{8,8} = \gamma\overline{\mathbb{E}\mathrm{tr}}(\hat{K}^{-1}) = G^0_{2,2} = \gamma\bar{\tau},$$
$$G^4_{3,6} = G^4_{9,12} = \frac{\gamma\sqrt{\rho_{\mathrm{s}}}\,\overline{\mathbb{E}\mathrm{tr}}[\Sigma_{\mathrm{s}} W^\top \hat{K}^{-1} W]}{N} = G^0_{3,6} = \gamma\sqrt{\rho_{\mathrm{s}}}\bar{\tau}\mathcal{I}^{\mathrm{s}}_{1,1},$$
$$G^4_{5,4} = G^4_{11,10} = -\frac{\sqrt{\rho_{\mathrm{s}}}\,\overline{\mathbb{E}\mathrm{tr}}[\Sigma_{\mathrm{s}} Z K^{-1} Z^\top]}{d} = G^0_{5,4} = -\frac{\sqrt{\rho_{\mathrm{s}}}\tau\mathcal{I}^{\mathrm{s}}_{1,1}}{\phi}.$$

Plugging these into (30), we have the following self-consistent equations

$$G^4_{2,8} = \frac{\sqrt{\rho_{\mathrm{s}}}\psi\phi^2 G^4_{5,10}}{(\phi + \rho_{\mathrm{s}}\tau\psi(\omega_{\mathrm{s}} + \mathcal{I}^{\mathrm{s}}_{1,1}))^2}, \quad G^4_{5,10} = -\frac{\rho_{\mathrm{s}}\tau^2}{\phi}\mathcal{I}^{\mathrm{s}}_{2,2} + \frac{\rho_{\mathrm{s}}^{\frac{3}{2}}\tau^2}{\phi}\mathcal{I}^{\mathrm{s}}_{2,2} G^4_{2,8}$$

Solving for $G^4_{2,8}$ and plugging in to D_{32}, we get

$$D_{32} = -\frac{2\rho_j\omega_j}{\sqrt{\rho_{\mathrm{s}}}} G^4_{2,8} = \frac{2\rho_j\omega_j\psi\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2}}{\rho_{\mathrm{s}}(\phi - \psi\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})}. \tag{31}$$

B.1.9. Computation of $\mathrm{Dis}^j_{\mathrm{SW}}(\phi, \psi, \gamma)$

Combining equations (26), (17), (27), (29), (31), we get

$$\mathrm{Dis}^j_{\mathrm{SW}}(\phi, \psi, \gamma) = \mathrm{Dis}^j_{\mathrm{I}}(\phi, \psi, \gamma) - \frac{2\rho_j\psi\kappa^2(\omega_j + \phi\mathcal{I}^j_{1,2})\mathcal{I}^{\mathrm{s}}_{2,2}}{\rho_{\mathrm{s}}(\phi - \psi\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})}.$$

B.2. Proof of Corollary 3.2

Since $\kappa \le 1/\omega_{\mathrm{s}}$ for any $\gamma > 0$ by (4), we know $\lim_{\gamma\to 0}\gamma\kappa = 0$. Thus from (6), we have

$$\lim_{\gamma\to 0}\gamma\tau = \frac{|\psi - \phi| + \psi - \phi}{2\psi}, \quad \lim_{\gamma\to 0}\gamma\bar{\tau} = 1 - \frac{\psi}{\phi} + \frac{\psi}{\phi}\lim_{\gamma\to 0}\gamma\tau = \frac{|\psi - \phi| + \phi - \psi}{2\phi}.$$

By Condition 2.3 and the dominated convergence theorem, the functionals $\mathcal{I}^{\rm s}_{a,b}, \mathcal{I}^{\rm t}_{a,b}$ and their derivatives with respect to κ are continuous in κ. Applying the implicit function theorem to the self-consistent equation (4), viewing it as a function of κ and γ, we find that κ is differentiable with respect to γ and thus continuous. Therefore, the limit of κ, $\mathcal{I}^{\rm s}_{a,b}$, $\mathcal{I}^{\rm t}_{a,b}$ when $\gamma \to 0$ is well defined. Plugging these limits into Theorem 3.1, we reach

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm I}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\rm s}|\phi-\psi|}(\sigma_\varepsilon^2+\mathcal{I}^{\rm s}_{1,1})(\omega_j+\mathcal{I}^j_{1,1}) + \begin{cases} \frac{2\rho_j\kappa(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})\mathcal{I}^j_{2,2}}{\rho_{\rm s}(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} & \phi>\psi, \\ \frac{2\rho_j\kappa(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\rm s}_{2,2}}{\rho_{\rm s}(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} & \phi<\psi, \end{cases}$$

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm SS}(\phi,\psi,\gamma) = \lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm I}(\phi,\psi,\gamma) - \frac{2\rho_j\kappa^2(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})\mathcal{I}^j_{2,2}}{\rho_{\rm s}(1-\kappa^2\mathcal{I}^{\rm s}_{2,2})}, \tag{32}$$

and

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm SW}(\phi,\psi,\gamma) = \lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm I}(\phi,\psi,\gamma) - \frac{2\rho_j\psi\kappa^2(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\rm s}_{2,2}}{\rho_{\rm s}(\phi-\psi\kappa^2\mathcal{I}^{\rm s}_{2,2})}. \tag{33}$$

From the equation (5), we have $\mathcal{I}^{\rm s}_{1,1} = \phi\mathcal{I}^{\rm s}_{1,2} + \kappa\mathcal{I}^{\rm s}_{2,2}$. Also by (4) and (6), $\omega_{\rm s} = \frac{1-\gamma\tau}{\kappa} - \mathcal{I}^{\rm s}_{1,1}$. Therefore,

$$\omega_{\rm s} + \phi\mathcal{I}^{\rm s}_{1,2} = \frac{1-\gamma\tau}{\kappa} - \mathcal{I}^{\rm s}_{1,1} + \phi\mathcal{I}^{\rm s}_{1,2} = \frac{1-\gamma\tau}{\kappa} - \kappa\mathcal{I}^{\rm s}_{2,2}. \tag{34}$$

In the ridgeless limit $\gamma \to 0$, the equation (34) gives

$$\lim_{\gamma\to 0} \frac{1}{\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2}} = \begin{cases} \lim_{\gamma\to 0} \frac{\kappa}{1-\kappa^2\mathcal{I}^{\rm s}_{2,2}} & \phi>\psi, \\ \lim_{\gamma\to 0} \frac{\psi\kappa}{\phi-\psi\kappa^2\mathcal{I}^{\rm s}_{2,2}} & \phi<\psi. \end{cases} \tag{35}$$

Putting (32), (33), (35) together, we conclude

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm SS}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\rm s}|\phi-\psi|}(\sigma_\varepsilon^2+\mathcal{I}^{\rm s}_{1,1})(\omega_j+\mathcal{I}^j_{1,1}) + \begin{cases} 0 & \phi>\psi, \\ \frac{2\rho_j\kappa}{\rho_{\rm s}}\left(\frac{(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\rm s}_{2,2}}{\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2}} - \frac{\kappa(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})\mathcal{I}^j_{2,2}}{1-\kappa^2\mathcal{I}^{\rm s}_{2,2}}\right) & \phi<\psi, \end{cases}$$

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm SW}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\rm s}|\phi-\psi|}(\sigma_\varepsilon^2+\mathcal{I}^{\rm s}_{1,1})(\omega_j+\mathcal{I}^j_{1,1}) + \begin{cases} \frac{2\rho_j\kappa}{\rho_{\rm s}}\left(\frac{(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})\mathcal{I}^j_{2,2}}{\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2}} - \frac{\psi\kappa(\omega_j+\phi\mathcal{I}^j_{1,2})\mathcal{I}^{\rm s}_{2,2}}{\phi-\psi\kappa^2\mathcal{I}^{\rm s}_{2,2}}\right) & \phi>\psi, \\ 0 & \phi<\psi. \end{cases}$$

B.3. Proof of Theorem 4.1

By Corollary 3.2, disagreement in the ridgeless and overparametrized regime is given by

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm I}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\rm s}|\phi-\psi|}(\sigma_\varepsilon^2+\mathcal{I}^{\rm s}_{1,1})(\omega_j+\mathcal{I}^j_{1,1}) + \frac{2\rho_j\kappa(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})\mathcal{I}^j_{2,2}}{\rho_{\rm s}(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})},$$

$$\lim_{\gamma\to 0} \mathrm{Dis}^j_{\rm SS}(\phi,\psi,\gamma) = \frac{2\rho_j\psi\kappa}{\rho_{\rm s}|\phi-\psi|}(\sigma_\varepsilon^2+\mathcal{I}^{\rm s}_{1,1})(\omega_j+\mathcal{I}^j_{1,1}).$$

The self-consistent equation (7) in the overpametrized regime $\phi > \psi$ is

$$\kappa = \frac{1}{\omega_{\rm s}+\mathcal{I}^{\rm s}_{1,1}(\kappa)},$$

which is independent of ψ. Consequently, the unique positive solution κ is also independent of ψ. This proves that the slope a and the intercept $b_{\rm I}$ defined in Theorem 4.1 are independent of ψ as well. Checking the equation (9) can be done by using (35) and a simple algebra.

B.4. Proof of Theorem 4.3

Let $a(\gamma), b_{\rm I}(\gamma), b_{\rm SS}(\gamma)$ be defined by (8), but with κ in the self-consistent equation (4) with general γ, instead of the self-consistent equation (7) in the ridgeless limit. With this notation, we have $a = a(0), b_{\rm I} = b_{\rm I}(0), b_{\rm SS} = b_{\rm SS}(0)$. By Theorem 3.1 and the triangle inequality, deviation from the line is bounded by

$$\begin{aligned}
&|\mathrm{Dis}_i^{\rm t}(\phi,\psi,\gamma) - a\mathrm{Dis}_i^{\rm s}(\phi,\psi,\gamma) - b_i| \\
&\qquad \le |\mathrm{Dis}_i^{\rm t}(\phi,\psi,\gamma) - a(\gamma)\mathrm{Dis}_i^{\rm s}(\phi,\psi,\gamma) - b_i(\gamma)| + |a(\gamma)-a(0)||\mathrm{Dis}_i^{\rm s}(\phi,\psi,\gamma)| + |b_i(\gamma)-b_i(0)| \\
&\qquad \le A_1 + A_2 + \mathrm{Dis}_i^{\rm s}(\phi,\psi,\gamma)|a(\gamma)-a(0)| + |b_i(\gamma)-b_i(0)|, \quad i \in \{\mathrm{I}, \mathrm{SS}\},
\end{aligned} \tag{36}$$

where

$$\begin{aligned}
A_1 &= \frac{2\psi\gamma\tau\kappa\mathcal{I}^{\rm s}_{2,2}|\rho_{\rm t}(\omega_{\rm t}+\phi\mathcal{I}^{\rm t}_{1,2}) - a\rho_{\rm s}(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})|}{\phi\gamma + \rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})}, \\
A_2 &= 2(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})|\rho_{\rm t}\mathcal{I}^{\rm t}_{2,2} - a\rho_{\rm s}\mathcal{I}^{\rm s}_{2,2}|\left|\frac{\kappa\phi\gamma\bar\tau}{\phi\gamma+\rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} - \frac{\kappa^2}{\rho_{\rm s}(1-\kappa^2\mathcal{I}^{\rm s}_{2,2})}\right|.
\end{aligned}$$

In what follows, we bound each of these terms. We will use $O(\cdot)$ notation to hide constants depending on $\phi, \mu, \sigma_\varepsilon^2, \sigma$. For example, we can write $\mathcal{I}^j_{a,b} = O(1)$ for $j \in \{\mathrm{s},\mathrm{t}\}$ since we assume in Condition 2.3 that μ is compactly supported.

B.4.1. Bounding A_1

We know $a \le \rho_{\rm t}(\omega_{\rm t}+\mathcal{I}^{\rm t}_{1,1})/\rho_{\rm s}\omega_{\rm s}$ by (8). Thus,

$$\mathcal{I}^{\rm s}_{2,2}|\rho_{\rm t}(\omega_{\rm t}+\phi\mathcal{I}^{\rm t}_{1,2}) - a\rho_{\rm s}(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})| = O(1). \tag{37}$$

By (6) and since $\sqrt{x^2+y^2} \le |x|+|y|$ for any $x, y \in \mathbb{R}$,

$$2\psi\gamma\tau = \sqrt{(\psi-\phi)^2 + 4\kappa\psi\phi\gamma/\rho_{\rm s}} + \psi - \phi \le \sqrt{\frac{4\kappa\psi\phi\gamma}{\rho_{\rm s}}} = O(\sqrt{\psi\gamma}). \tag{38}$$

Again by (6), $\psi\gamma\tau + \phi\gamma\bar\tau = \sqrt{(\psi-\phi)^2 + 4\kappa\psi\phi\gamma/\rho_{\rm s}}$. Therefore,

$$\frac{\kappa}{\phi\gamma+\rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} \le \frac{\kappa}{\rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} = O\left(\frac{1}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{39}$$

Here, we used $\kappa \le \frac{1}{\omega_{\rm s}} = O(1)$ by (4). Combining (37), (38), (39), we reach

$$A_1 = O\left(\frac{\sqrt{\psi\gamma}}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{40}$$

B.4.2. Bounding A_2

Similar to (37), we have

$$2(\sigma_\varepsilon^2+\phi\mathcal{I}^{\rm s}_{1,2})|\rho_{\rm t}\mathcal{I}^{\rm t}_{2,2} - a\rho_{\rm s}\mathcal{I}^{\rm s}_{2,2}| = O(1). \tag{41}$$

By (34),

$$\frac{\kappa^2}{\rho_{\rm s}(1-\kappa^2\mathcal{I}^{\rm s}_{2,2})} = \frac{\kappa^2}{\rho_{\rm s}[\gamma\tau+\kappa(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})]}. \tag{42}$$

From (42) and $\kappa = \gamma\rho_{\rm s}\tau\bar\tau$,

$$\begin{aligned}
&\left|\frac{\kappa\phi\gamma\bar\tau}{\phi\gamma+\rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})} - \frac{\kappa^2}{\rho_{\rm s}(1-\kappa^2\mathcal{I}^{\rm s}_{2,2})}\right| \\
&= \frac{\kappa^2(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})\psi\gamma\tau}{[\phi\gamma+\rho_{\rm s}(\psi\gamma\tau+\phi\gamma\bar\tau)(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})][\gamma\tau+\kappa(\omega_{\rm s}+\phi\mathcal{I}^{\rm s}_{1,2})]}.
\end{aligned}$$

From (38), (39), and $\kappa(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})/[\gamma\tau+\kappa(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})]\leq 1$, we get

$$\left|\frac{\kappa\phi\gamma\bar{\tau}}{\phi\gamma+\rho_{\mathrm{s}}(\psi\gamma\tau+\phi\gamma\bar{\tau})(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})}-\frac{\kappa^2}{\rho_{\mathrm{s}}(1-\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2})}\right|=O\left(\frac{\sqrt{\psi\gamma}}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{43}$$

Putting (41) and (43) together,

$$A_2=O\left(\frac{\sqrt{\psi\gamma}}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{44}$$

B.4.3. Bounding $\mathrm{Dis}^{\mathrm{s}}_{\mathrm{I}}(\phi,\psi,\gamma)$ and $\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SS}}(\phi,\psi,\gamma)$

By Theorem 3.1 and the equations (6), (38), (39), we have

$$\mathrm{Dis}^{\mathrm{s}}_{\mathrm{I}}(\phi,\psi,\gamma)=O\left(\frac{1+\sqrt{\psi\gamma}}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{45}$$

By Theorem 3.1 and the equations (38), (39), (43), we have

$$\mathrm{Dis}^{\mathrm{s}}_{\mathrm{SS}}(\phi,\psi,\gamma)=O\left(\frac{\psi+\sqrt{\psi\gamma}}{1-\psi/\phi+\sqrt{\psi\gamma}}\right). \tag{46}$$

B.4.4. Bounding $|a(\gamma)-a(0)|$

From the argument in Section B.2, we know $a(\gamma)$ is differentiable with respect to γ. By the chain rule and (5),

$$\frac{\partial a}{\partial\gamma}=\frac{\partial\kappa}{\partial\gamma}\times\frac{-\mathcal{I}^{\mathrm{t}}_{2,2}(\omega_{\mathrm{s}}+\mathcal{I}^{\mathrm{s}}_{1,1})+\mathcal{I}^{\mathrm{s}}_{2,2}(\omega_{\mathrm{t}}+\mathcal{I}^{\mathrm{t}}_{1,1})}{(\omega_{\mathrm{s}}+\mathcal{I}^{\mathrm{s}}_{1,1})^2}. \tag{47}$$

By implicit differentiation of (4), we have

$$\frac{\partial\kappa}{\partial\gamma}=-\frac{\kappa}{\phi\gamma+\rho_{\mathrm{s}}(\psi\gamma\tau+\phi\gamma\bar{\tau})(\omega_{\mathrm{s}}+\phi\mathcal{I}^{\mathrm{s}}_{1,2})}. \tag{48}$$

We have $|(-\mathcal{I}^{\mathrm{t}}_{2,2}(\omega_{\mathrm{s}}+\mathcal{I}^{\mathrm{s}}_{1,1})+\mathcal{I}^{\mathrm{s}}_{2,2}(\omega_{\mathrm{t}}+\mathcal{I}^{\mathrm{t}}_{1,1}))/(\omega_{\mathrm{s}}+\mathcal{I}^{\mathrm{s}}_{1,1})^2|=O(1)$ and

$$\left|\frac{\partial\kappa}{\partial\gamma}\right|=O\left(\frac{1}{\sqrt{(\psi-\phi)^2+\psi\phi\gamma}}\right)$$

since $\psi\gamma\tau+\phi\gamma\bar{\tau}=\sqrt{(\psi-\phi)^2+4\kappa\psi\phi\gamma/\rho_{\mathrm{s}}}$. Therefore,

$$\begin{aligned}|a(\gamma)-a(0)|&=\left|\int_0^\gamma\frac{\partial a}{\partial\gamma}(u)du\right|\leq\int_0^\gamma\left|\frac{\partial a}{\partial\gamma}(u)\right|du\\&=O\left(\int_0^\gamma\frac{1}{\sqrt{(\psi-\phi)^2+\psi\phi u}}du\right)=O\left(\frac{\gamma}{1-\psi/\phi+\sqrt{\psi\gamma}}\right).\end{aligned} \tag{49}$$

B.4.5. Bounding $|b_{\mathrm{I}}(\gamma)-b_{\mathrm{I}}(0)|$

From the argument in Section B.2, we know $b_{\mathrm{I}}(\gamma)$ is differentiable with respect to γ. In (8), the terms $\frac{\kappa^2}{1-\kappa^2\mathcal{I}^{\mathrm{s}}_{2,2}}$, $\sigma_\varepsilon^2+\phi\mathcal{I}^{\mathrm{s}}_{1,2}$, $\rho_{\mathrm{t}}-a\rho_{\mathrm{s}}\mathcal{I}^{\mathrm{s}}_{2,2}$ and their derivatives with respect to κ are $O(1)$. Thus,

$$\left|\frac{\partial b_{\mathrm{I}}}{\partial\gamma}\right|=O\left(\left|\frac{\partial\kappa}{\partial\gamma}\right|\right)=O\left(\frac{1}{\sqrt{(\psi-\phi)^2+\psi\phi\gamma}}\right).$$

Therefore,

$$|b_{\mathrm{I}}(\gamma) - b_{\mathrm{I}}(0)| = \left| \int_0^\gamma \frac{\partial b_{\mathrm{I}}}{\partial \gamma}(u) du \right| \le \int_0^\gamma \left| \frac{\partial b_{\mathrm{I}}}{\partial \gamma}(u) \right| du$$
$$= O\left(\int_0^\gamma \frac{1}{\sqrt{(\psi-\phi)^2 + \psi\phi u}} du \right) = O\left(\frac{\gamma}{1 - \psi/\phi + \sqrt{\psi\gamma}} \right). \tag{50}$$

Theorem 4.3 is proved by combining the equations (36), (40), (44), (45), (46), (49), (50).

B.5. Proof of Corollary 4.4

By $E_j = B_j + V_j = B_j + \frac{1}{2}\mathrm{Dis}_{\mathrm{I}}^j(\phi, \psi, \gamma)$ and (51), we have

$$|E_{\mathrm{t}} - aE_{\mathrm{s}} - b_{\mathrm{risk}}| \le \frac{1}{2}|\mathrm{Dis}_{\mathrm{I}}^{\mathrm{t}}(\phi, \psi, \gamma) - a\mathrm{Dis}_{\mathrm{I}}^{\mathrm{s}}(\phi, \psi, \gamma) - b_{\mathrm{I}}| + \left| B_{\mathrm{t}} - aB_{\mathrm{s}} - \lim_{\gamma \to 0}(B_{\mathrm{t}} - aB_{\mathrm{s}}) \right|.$$

Since the derivatives of $\mathcal{I}_{1,1}^j, \mathcal{I}_{1,2}^j$ with respect to γ is $O(1)$. We have

$$\left| B_{\mathrm{t}} - aB_{\mathrm{s}} - \lim_{\gamma \to 0}(B_{\mathrm{t}} - aB_{\mathrm{s}}) \right| = O(\gamma)$$

by the mean value theorem. The conclusion follows from Theorem 4.3.

C. Recap of Tripuraneni et al. (2021)

In this section, we restate some relevant results of Tripuraneni et al. (2021), in the special cases $\Sigma^* = \Sigma_{\mathrm{s}}$ or $\Sigma^* = \Sigma_{\mathrm{t}}$. See Tripuraneni et al. (2021) for the original theorems.

For a test distribution $x \sim \mathsf{N}(0, \Sigma^*)$, define the risk by

$$E_{\Sigma^*} = \mathbb{E}_{x,\beta,X,Y,W}[(\beta^\top x - \hat{y}_{W,X,Y}(x))^2].$$

We have the following bias-variance decomposition

$$E_{\Sigma^*} = \mathbb{E}_{x,\beta}[(\beta^\top x - \mathbb{E}_{W,X,Y}[\hat{y}_{W,X,Y}(x)])^2] + \mathbb{E}_{x,\beta}[\mathbb{V}_{W,X,Y}(\hat{y}_{W,X,Y}(x))]$$
$$= B_{\Sigma^*} + V_{\Sigma^*}.$$

We consider the high-dimensional limit $n, d, N \to \infty$ with $d/n \to \phi$ and $d/N \to \psi$ of the above quantities when $\Sigma^* = \Sigma_{\mathrm{s}}$ or $\Sigma^* = \Sigma_{\mathrm{t}}$,

$$E_j = \lim_{n,d,N\to\infty} E_{\Sigma_j}, \quad B_j = \lim_{n,d,N\to\infty} B_{\Sigma_j}, \quad V_j = \lim_{n,d,N\to\infty} V_{\Sigma_j}, \quad j \in \{\mathrm{s}, \mathrm{t}\}.$$

Theorem C.1 (Theorem 5.1 of Tripuraneni et al. (2021))**.** *For $j \in \{\mathrm{s}, \mathrm{t}\}$, the asymptotic bias and variance are given by*

$$B_j = \left(1 - \sqrt{\frac{\rho_j}{\rho_{\mathrm{s}}}}\right)^2 m_j + 2\left(1 - \sqrt{\frac{\rho_j}{\rho_{\mathrm{s}}}}\right)\sqrt{\frac{\rho_j}{\rho_{\mathrm{s}}}}\mathcal{I}_{1,1}^j + \frac{\rho_j\phi}{\rho_{\mathrm{s}}}\mathcal{I}_{1,2}^j,$$
$$V_j = -\frac{\rho_j\psi}{\phi}\frac{\partial\kappa}{\partial\gamma}\left[\mathcal{I}_{1,1}^{\mathrm{s}}(\omega_{\mathrm{s}} + \phi\mathcal{I}_{1,2}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^j) + \frac{\phi^2}{\psi}\gamma\bar{\tau}\mathcal{I}_{1,2}^{\mathrm{s}}\mathcal{I}_{2,2}^j\right.$$
$$\left. + \gamma\tau\mathcal{I}_{2,2}^{\mathrm{s}}(\omega_j + \phi\mathcal{I}_{1,2}^j) + \sigma_\varepsilon^2\left((\omega_{\mathrm{s}} + \phi\mathcal{I}_{1,2}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^j) + \frac{\phi}{\psi}\gamma\bar{\tau}\mathcal{I}_{2,2}^j\right)\right],$$

where $\kappa, \tau, \bar{\tau}$ are defined in (4) and (6).

In the ridgeless limit $\gamma \to 0$, the variance V_j is further simplified as follows.

Corollary C.2 (Corollary 5.1 of Tripuraneni et al. (2021)). *For $j \in \{\mathrm{s}, \mathrm{t}\}$, the asymptotic variance in the ridgeless limit is*

$$\lim_{\gamma \to 0} V_j = \frac{\rho_j \psi \kappa}{\rho_{\mathrm{s}} |\phi - \psi|} (\sigma_\varepsilon^2 + \mathcal{I}_{1,1}^{\mathrm{s}})(\omega_j + \mathcal{I}_{1,1}^j) + \begin{cases} \frac{\rho_j \kappa}{\rho_{\mathrm{s}}} \left(1 - \frac{\kappa(\omega_{\mathrm{s}} - \sigma_\varepsilon^2)}{1 - \kappa^2 \mathcal{I}_{2,2}^{\mathrm{s}}}\right) \mathcal{I}_{2,2}^j & \phi \geq \psi, \\ \frac{\rho_j \kappa^2 \psi \mathcal{I}_{2,2}^{\mathrm{s}}}{\rho_{\mathrm{s}}(\phi - \kappa^2 \psi \mathcal{I}_{2,2}^{\mathrm{s}})} (\omega_j + \phi \mathcal{I}_{1,2}^j) & \phi < \psi, \end{cases}$$

where κ is defined in (7).

Another important observation is that there is a linear relation between the asymptotic error under the source and target domain.

Proposition C.3 (Proposition 5.6 of Tripuraneni et al. (2021)). *We assume ϕ is fixed. In the ridgeless limit $\gamma \to 0$ and the overparametrized regime $\phi \geq \psi$, the error E_{t} is linear in E_{s}, as a function of ψ. That is,*

$$\lim_{\gamma \to 0} E_{\mathrm{t}} = b_{\text{risk}} + \frac{\rho_{\mathrm{t}}(\omega_{\mathrm{t}} + \mathcal{I}_{1,1}^{\mathrm{t}})}{\rho_{\mathrm{s}}(\omega_{\mathrm{s}} + \mathcal{I}_{1,1}^{\mathrm{s}})} \lim_{\gamma \to 0} E_{\mathrm{s}},$$

where the intercept

$$b_{\text{risk}} = \frac{1}{2} b_{\mathrm{I}} + \lim_{\gamma \to 0} (B_{\mathrm{t}} - a B_{\mathrm{s}}) \tag{51}$$

and the slope $\rho_{\mathrm{t}}(\omega_{\mathrm{t}} + \mathcal{I}_{1,1}^{\mathrm{t}})/\rho_{\mathrm{s}}(\omega_{\mathrm{s}} + \mathcal{I}_{1,1}^{\mathrm{s}})$ are independent of ψ.

D. Additional Experiments

D.1. Estimation of the Slope

Let $\hat{\Sigma}_{\mathrm{s}}, \hat{\Sigma}_{\mathrm{t}}$ be sample covariance of test inputs from the source and target domain, respectively. Denote the eigenvalues and corresponding eigenvectors of $\hat{\Sigma}_{\mathrm{s}}$ by $\hat{\lambda}_1^{\mathrm{s}}, \ldots, \hat{\lambda}_d^{\mathrm{s}}$ and $\hat{v}_1, \ldots, \hat{v}_d$. Define $\hat{\lambda}_i^{\mathrm{t}} = \hat{v}_i^\top \hat{\Sigma}_{\mathrm{t}} \hat{v}_i$ for $i \in [d]$. For $j \in \{\mathrm{s}, \mathrm{t}\}$, we estimate $\mathcal{I}_{a,b}^j(\kappa)$ by

$$\hat{\mathcal{I}}_{a,b}^j(\kappa) = \frac{\phi}{d} \sum_{i=1}^d \frac{(\hat{\lambda}_i^{\mathrm{s}})^{a-1} \hat{\lambda}_i^j}{(\phi + \kappa \hat{\lambda}_i^{\mathrm{s}})^b}.$$

We estimate the constants defined in (3) by replacing m_j with $\hat{m}_j = \overline{\mathrm{tr}}(\hat{\Sigma}_j)$, $j \in \{\mathrm{s}, \mathrm{t}\}$. Now, the self-consistent equation (7) is estimated by

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

and its unique non-negative solution is denoted by $\hat{\kappa}$. The existence and uniqueness of $\hat{\kappa}$ follows from Lemma A1.2 of Tripuraneni et al. (2021). We use

$$\hat{a} = \frac{\hat{\rho}_t(\hat{\omega}_{\mathrm{t}} + \hat{\mathcal{I}}_{1,1}^{\mathrm{t}}(\hat{\kappa}))}{\hat{\rho}_s(\hat{\omega}_{\mathrm{s}} + \hat{\mathcal{I}}_{1,1}^{\mathrm{s}}(\hat{\kappa}))}$$

as an estimate of the slope $a = \rho_{\mathrm{t}}(\omega_{\mathrm{t}} + \mathcal{I}_{1,1}^{\mathrm{t}})/\rho_{\mathrm{s}}(\omega_{\mathrm{s}} + \mathcal{I}_{1,1}^{\mathrm{s}})$.

D.2. Deviation from the Line

Figure 5 displays deviation from the line for I disagreement and risk, when non-zero ridge regularization γ is used. Similar to Figure 3 **(b)**, the deviation is smaller for γ closer to zero. However, unlike SS disagreement, the deviation is non-zero even in the infinite overparameterization limit $\psi \to 0$. This is consistent with the upper bound we present in Theorem 4.3 and Corollary 4.4.

D.3. Varying Corruption Severity

CIFAR-10-C and Tiny ImageNet-C have different severity of corruption ranging from 1 to 5. We only included a few selected results in the main text due to space limitations. We present the plots for all severity levels in Figure 7.

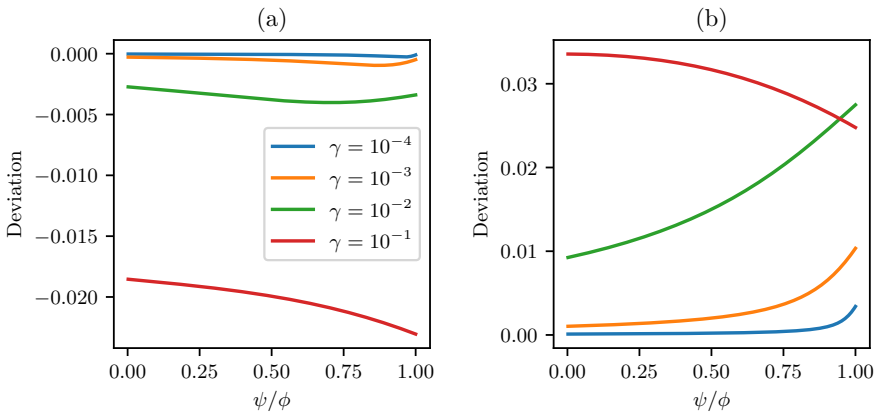

Figure 5. **(a)** Deviation from the line, $\text{Dis}_{\text{I}}^{\text{t}}(\phi, \psi, \gamma) - a\text{Dis}_{\text{I}}^{\text{s}}(\phi, \psi, \gamma) - b_I$, as a function of ψ for non-zero γ. **(b)** Deviation from the line, $E_{\text{t}} - aE_{\text{s}} - b_{\text{risk}}$, as a function of ψ for non-zero γ. We use $\phi = 0.5$, $\sigma_\varepsilon^2 = 10^{-4}$, ReLU activation σ, and $\mu = 0.4\delta_{(0.1,1)} + 0.6\delta_{(1,0.1)}$

D.4. I and SW disagreement

In Figure 8, Figure 9, Figure 6 **(a), (b)**, we repeat the experiment in Section D.3 for I and SW disagreement. Since our theory suggests that the disagreement-on-the-line phenomenon does not occur for SW disagreement, we do not plot theoretical predictions for SW disagreement.

D.5. Accuracy and Agreement

In the main text, we consider disagreement and risk defined in terms of mean squared error, but here we present classification accuracy and 0-1 agreement as studied in Hacohen et al. (2020); Chen et al. (2021); Jiang et al. (2021); Nakkiran & Bansal (2020); Baek et al. (2022); Atanov et al. (2022); Pliushch et al. (2022); Kirsch & Gal (2022). See Figures 10 and Figure 6 **(c)**.

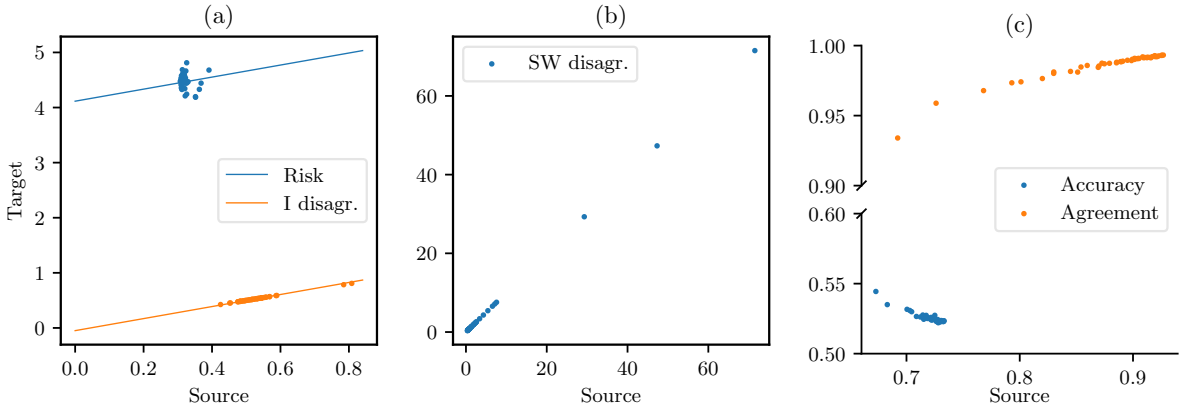

Figure 6. **(a)** Target vs. source independent disagreement of random features model trained on Camelyon17. **(b)** Target vs. source shared-weight disagreement of random features model trained on Camelyon17. **(c)** Target vs. source accuracy and agreement of random features model trained on Camelyon17; Experimental setting is identical to Section 5.

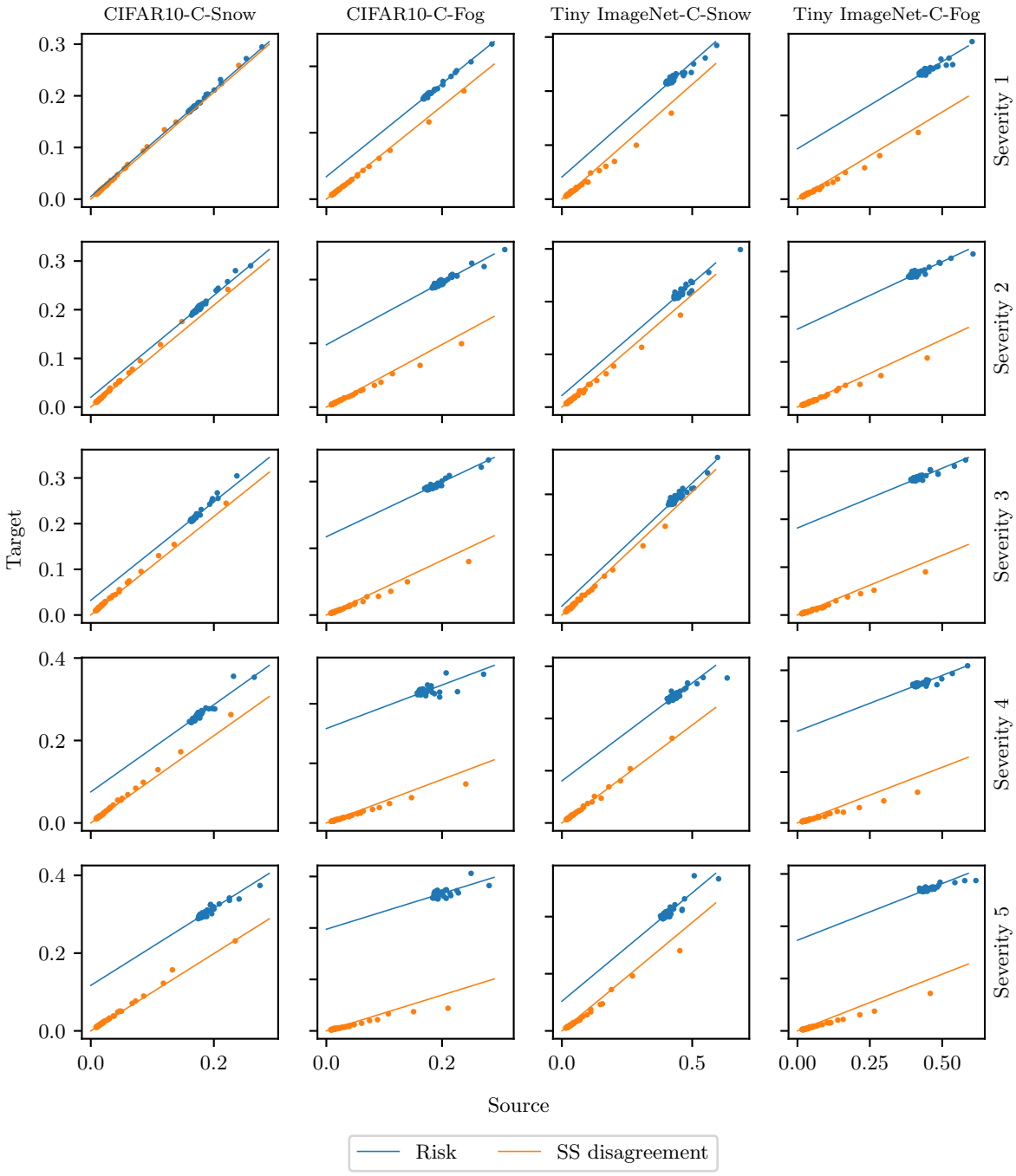

Figure 7. Target vs. source shared-sample disagreeement on CIFAR-10 and Tiny ImageNet with varying corruption severity. Experimental setting is identical to Section 5.

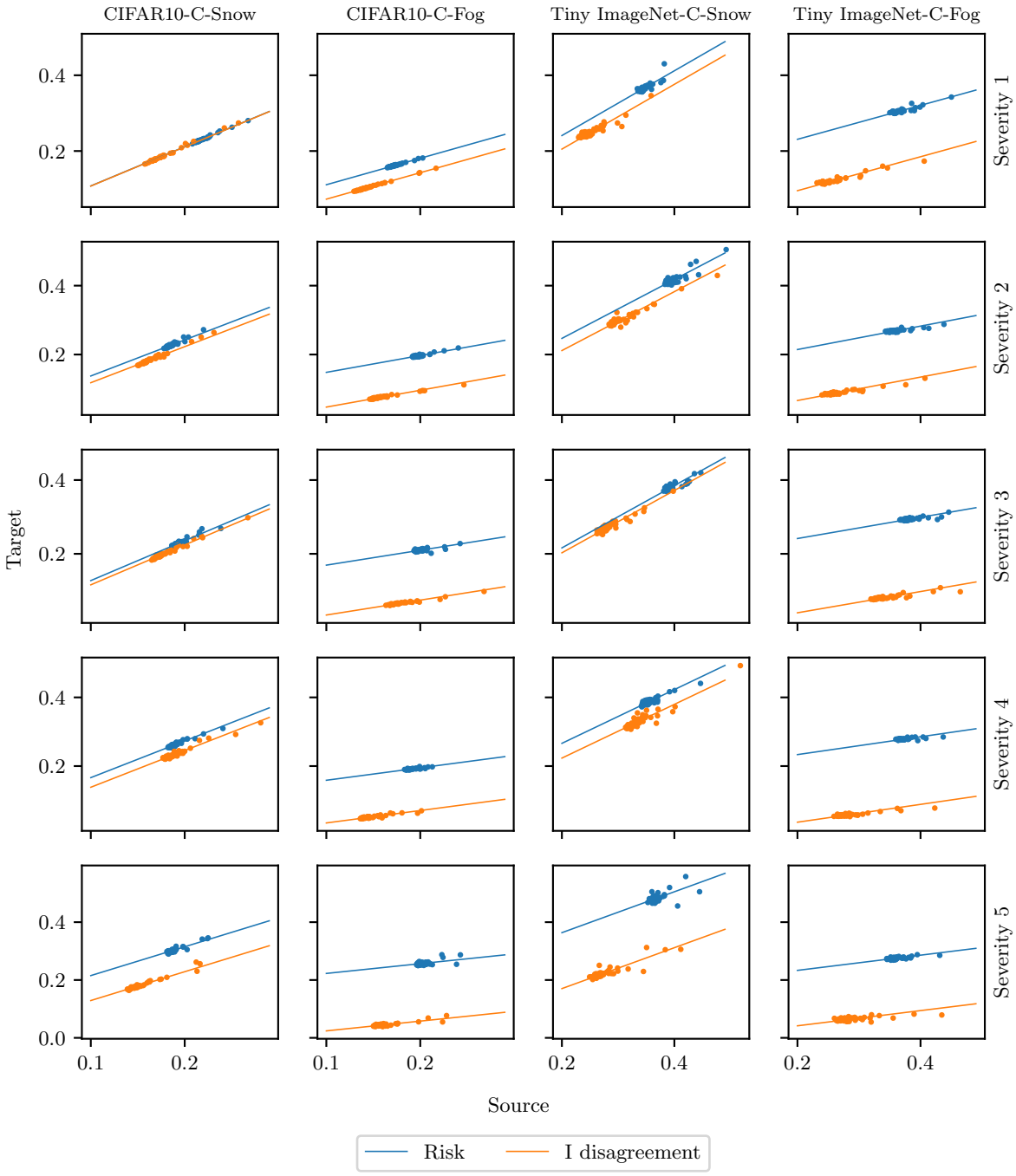

Figure 8. Target vs. source independent disagreeement on CIFAR-10 and Tiny ImageNet with varying corruption severity. Experimental setting is identical to Section 5.

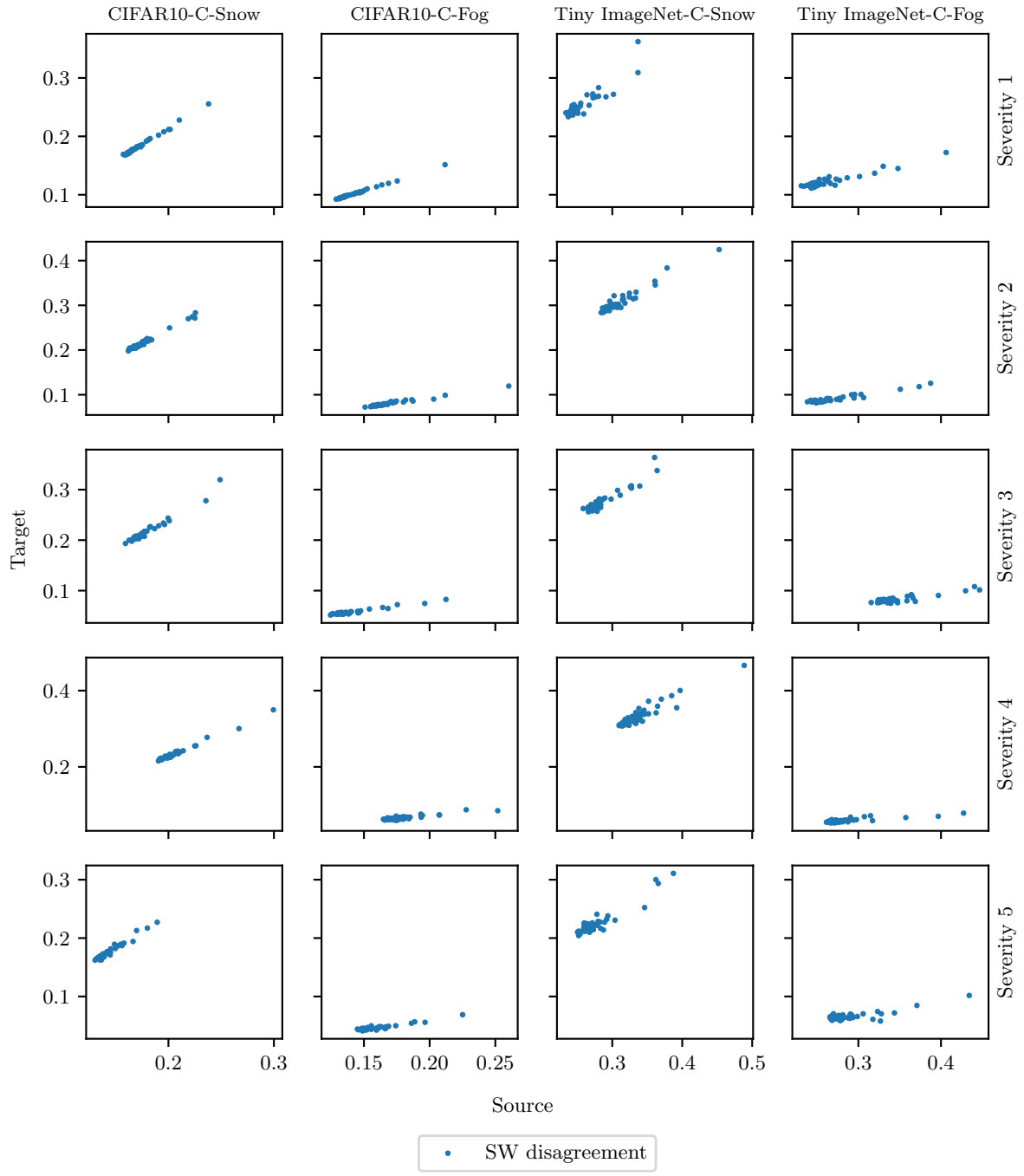

Figure 9. Target vs. source shared-weight disagreeement on CIFAR-10 and Tiny ImageNet with varying corruption severity. Experimental setting is identical to Section 5.

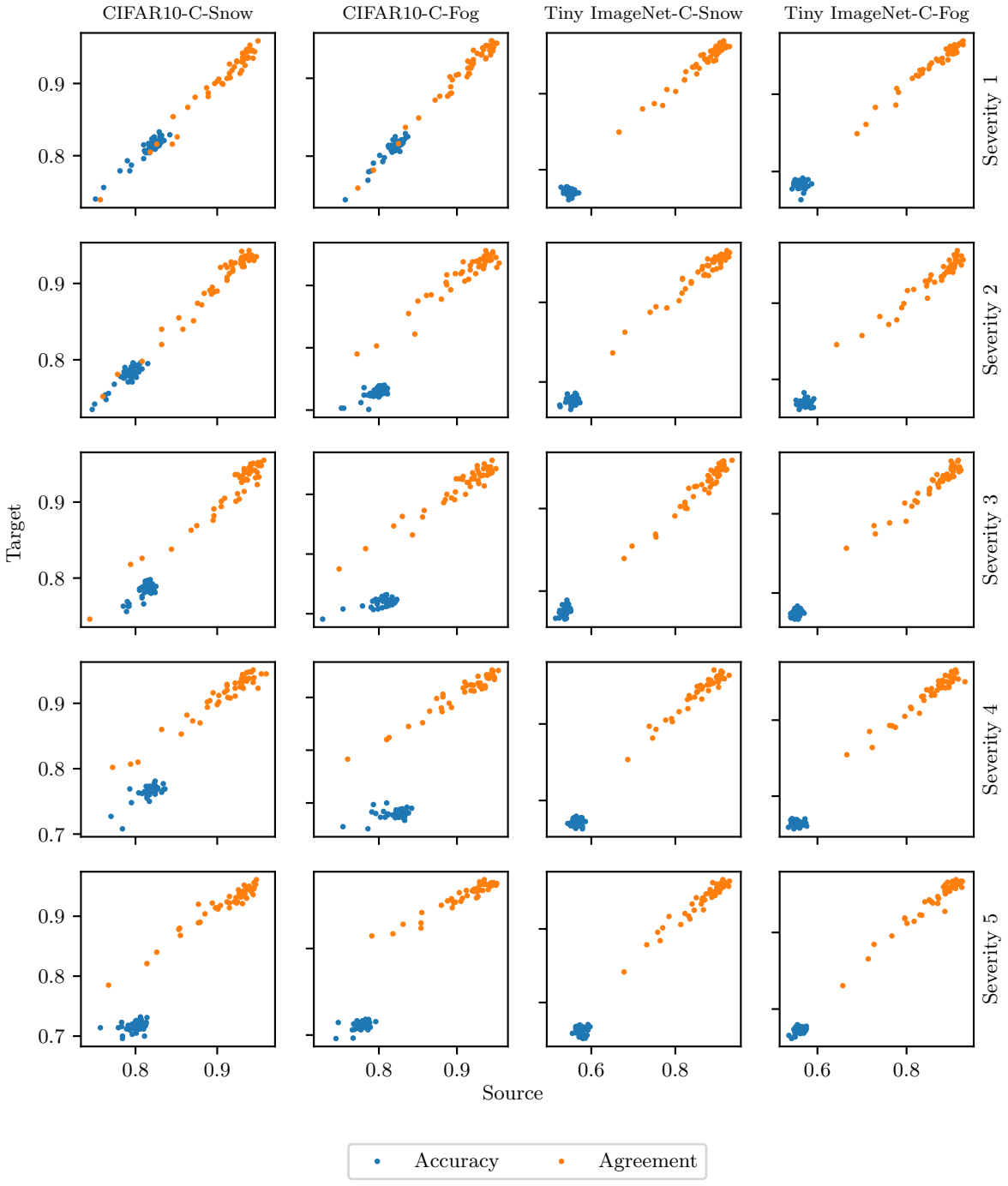

Figure 10. Target vs. source classification accuracy and agreement on CIFAR-10 and Tiny ImageNet with varying corruption severity. Experimental setting is identical to Section 5.