Characterizing the Generalization Error of Random Feature Regression with Arbitrary Data-Augmentation

Lucas Morisset[*]
Qube Research & Technologies
École Polytechnique

Alain Durmus
École Polytechnique

Adrien Hardy
Qube Research & Technologies

Abstract

This paper aims at analyzing the regularization effect that data augmentation induces on supervised regression methods in the proportional regime, where the number of covariates grows proportionally to the number of samples. We provide a tight characterization of the test error, measured in mean squared error, in terms only of the population quantities of the true data, as well as first and second order statistics of the augmentation scheme. Our results are valid under misspecified feature maps, and for any network architecture where only the last readout layer is trained, and the rest of the network is either frozen or randomly initialized. We specify our results in the case of Gaussian data, and show that our asymptotic characterization is tight in this setting.

1 Introduction

Data augmentation (DA) is now a standard ingredient in modern machine learning pipelines, with extensive empirical evidence reporting improvements in generalization across modalities and tasks [MM22b, WWL+25]. It is often used to encode task-relevant symmetries directly into the training procedure, for instance by encouraging invariance to image rotations or other transformations of the input [SK19, CDL20]. It has also been identified as one of the most effective regularization techniques across both supervised learning settings [Bis95, CZM+19, MM22b, WWL+25] and self-supervised/unsupervised learning [FGW+21, VAIB+25]. Domain-specific augmentation pipelines have been central to progress in computer vision [SK19, KBMB24], natural language processing [FGW+21, SKF21, BKR22], and time-series or audio applications [WSY+20, IU21, ITGP+23]. Despite these empirical successes, the benefits of DA remain highly task- and data-dependent, and augmentation schemes are often engineered in an ad hoc manner [FSTF16, CZM+19, LKK+19, HZYN20].

In contrast with this rich empirical literature, comprehensive theoretical analyses of DA remain relatively scarce. Two classical starting points are, first, the interpretation of additive Gaussian noise as a form of explicit (ridge-like) regularization [Bis95, LKDM24], and second, the idea that leveraging distributional invariances and group structure in the learning objective helps decrease the variance of the model without increasing its bias [CDL20]. Yet, when applied to modern and complex augmentation schemes, these works either provide only upper bounds on the generalization error [LKDM24], or require very strong assumptions on the data distribution (e.g. approximate invariance under the augmentation scheme [CDL20]). The present manuscript is motivated by the need for a tractable theoretical framework capable of capturing augmentation schemes used in deep learning, including those that may induce strong dependence across observations without preserving the data distribution.

Among the tractable proxy models used to study neural networks, random feature networks, obtained by training only the last layer of a random representation, have emerged as particularly natural

[*]Corresponding author: lucas.morisset@polytechnique.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

candidates [BHMM19, DL20, GLR+22, HL23]. Beyond their mathematical convenience, random networks at initialisation have proven to be a valuable theoretical testbed, and can capture empirically observed phenomena such as double-descent [MM22a] and benign overfitting [BLLT20]. As a result, tractable proxies for networks at initialisation, such as random feature models [BHMM19, DL20, GLR+22, HL23, SCDL23, SDCL24], have received considerable attention. Remarkably, it has also been observed that structured random networks, whose weights are sampled to match certain statistics of trained models, can achieve performance comparable to that of fully trained networks [GMRM23]. In some cases, matching second-order statistics alone is sufficient, which further supports random feature models as a natural testbed for deep learning. This perspective has led to a comprehensive theory of random feature models [SCDL23, ZVP23, BPH23, SDCL24]. Yet this theory has so far remained confined to the i.i.d. setting, and a central technical challenge of the present work is to extend it to the dependent regime induced by data augmentation.

More precisely, we aim to provide a precise asymptotic characterization of the out-of-sample error of feature regression trained with data augmentation. To this end, we consider a general augmentation framework that covers a broad class of modern augmentation schemes. Our results apply to any class of architecture with a trained linear readout and either frozen or randomly initialized representation, and they reveal that data augmentation can induce subtle and sometimes competing effects on generalization. Our contributions are as follows:

- We derive deterministic equivalents for the generalization error of random-feature regression trained with data augmentation under possible misspecification, extending previous results from [DL20, GLR+22, HL23, SCDL23, SDCL24] to the dependent setting. In particular, we obtain a closed-form deterministic quantity that approximates the true generalization error in the proportional regime, *i.e.*, when the number of features scales proportionally with the sample size. We also derive explicit non-asymptotic bounds for the corresponding approximation error as a function of the ridge parameter and the sample size.
- Beyond the test error, we derive deterministic equivalents for the resulting estimator, as well as for key quadratic forms and transformations arising in the generalization analysis, which we believe are of independent interest.
- We empirically validate the accuracy of the deterministic equivalents on synthetic Gaussian experiments across a range of dimensions, aspect ratios, augmentation strengths, and ridge parameters. We also investigate how data augmentation affects the bias and variance components of the error when the feature model is misspecified. One of our main findings is that augmentation can lead to a substantial reduction in variance and improved generalization over a broad range of ridge parameters, without necessarily inducing a monotonic increase in bias, contrary to the usual intuition drawn from correctly specified settings.

1.1 Notation

Motivated by high-dimensional statistics (and while all our results are non-asymptotic), we view the sample size n, the covariate dimenson $d = d(n)$, and the feature dimension $p = p(n)$ as variable scaling parameters. For notational simplicity, we typically omit the explicit dependence on n for non-scalar quantities, as it is always implicitly determined by the dimensions of the relevant vectors and matrices. Throughout, c denotes a universal constant, independent of n, d, p, and λ; its value may change from line to line. For a matrix $\mathbf{A}$, we write $\|\mathbf{A}\|_{\mathrm{F}}$ and $\|\mathbf{A}\|_{\mathrm{op}}$ for its Frobenius and operator norms, respectively. Moreover, for any $\mathbf{A} \in \mathbb{R}^{d\times n}$, any $i \in 1, \dots, n$, and any function $f : \mathbb{R}^d \to \mathbb{R}^p$, we denote by $\mathbf{A}_i$ the i-th column of $\mathbf{A}$ and define $f(\mathbf{A}) = (f(\mathbf{A}_1), \dots, f(\mathbf{A}_n)) \in \mathbb{R}^{p\times n}$. Finally, for a random matrix A, we define its variance by $\mathrm{Var}(A) = \mathbb{E}\left[\|A - \mathbb{E}[A]\|_{\mathrm{F}}^2\right]$, which corresponds to the sum of the entrywise variances. The matrix I_p denotes the $p \times p$ identity matrix. For two random matrices $A_1 \in \mathbb{R}^{p\times k}$ and $A_2 \in \mathbb{R}^{l\times k}$, we define the associated covariance matrix as $\mathrm{Cov}(A_1, A_2) = \mathbb{E}\left[(A_1 - \mathbb{E}[A_1])(A_2 - \mathbb{E}[A_2])^\top\right]$ and $\mathrm{Cov}(A_1) = \mathrm{Cov}(A_1, A_1)$. We denote by $\mathrm{N}(\mathbf{m}, \mathbf{\Sigma})$ the Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{\Sigma}$.

2 Settings

Consider a training dataset $\mathtt{Z} = (\mathtt{Z}_i)_{i=1}^n = \{(\mathtt{X}_i, \mathtt{Y}_i)\}_{i=1}^n$, where $\mathtt{X}_i \in \mathbb{R}^d$ and $\mathtt{Y}_i \in \mathbb{R}$ denote the covariates and responses respectively, satisfying the following condition.

H1. *We assume that the true covariates* $\mathtt{Z} = \{(\mathtt{X}_i, \mathtt{Y}_i)\}_{i=1}^n$ *are centered i.i.d. random variables, such that the responses write for some* $\varphi_\star : \mathbb{R}^d \to \mathbb{R}^{p_\star}$ *and* $\theta_\star \in \mathbb{R}^{p_\star}$, $p_\star \geqslant 1$,

$$\mathtt{Y}_i = \theta_\star^\top \varphi_\star(\mathtt{X}_i) + \varepsilon_i \ , \tag{1}$$

for all $i \leqslant n$, *and where* ε_i *is a centered, sub-Gaussian random variable with parameter* σ^2.

Let $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ be a feature map which may differ from $\varphi_\star$; note that we allow different codimensions p and $p_\star$. In particular, the learned representation $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ used by the estimator need not coincide with the ground-truth representation $\varphi_\star : \mathbb{R}^d \to \mathbb{R}^{p_\star}$ in (1), allowing for model misspecification. Equipped with φ, feature regression model aims to fit the data finding $\theta \in \mathbb{R}^p$ that minimizes the empirical risk function $\mathcal{L}(\theta) = n^{-1} \sum_{i=1}^n \big\{\mathtt{Y}_i - \theta^\top \varphi(\mathtt{X}_i)\big\}^2$. To improve performance various DA strategies have been proposed. In this paper, we model a data augmentation scheme as a random transformation $(\tau_\mathtt{x}, \tau_\mathtt{y})$ acting on feature/label pair. More precisely we consider $\tau_\mathtt{x} : \mathbb{R}^d \times \mathbb{R} \times \mathsf{E} \to \mathbb{R}^d$ and $\tau_\mathtt{y} : \mathbb{R}^d \times \mathbb{R} \times \mathsf{E} \to \mathbb{R}$, referred to as the feature and label transformations respectively, and where $(\mathsf{E}, \mathcal{E})$ is a measurable space. Then, a data augmentation scheme is completed by specifying a distribution $\mathcal{T}$ over $(\mathsf{E}, \mathcal{E})$. Artificial samples are then defined as $(\tau_\mathtt{x}(\mathtt{Z}_i, \eta), \tau_\mathtt{y}(\mathtt{Z}_i, \eta))$, where η is distributed according to $\mathcal{T}$ and is independent of the dataset $\mathtt{Z}$. This framework covers both input-only and label-dependent augmentations. Based on $(\tau_\mathtt{x}, \tau_\mathtt{y}, \mathcal{T})$, we can then consider the augmented risk functions defined as

$$\mathcal{L}_{\mathrm{Aug}}(\theta) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[\big\{ \tau_\mathtt{y}(\mathtt{Z}_i, \eta) - \theta^\top \varphi(\tau_\mathtt{x}(\mathtt{Z}_i, \eta)) \big\}^2 \ \Big| \ \mathtt{Z} \Big] \ , \tag{2}$$

and consider the estimator obtained by minimizing a linear combinaison of the two previous risk functions: for $\lambda \geqslant 0$ and $\alpha \in [0, 1]$,

$$\hat{\theta}_{\alpha,\lambda} := \arg\min_{\theta \in \mathbb{R}^p} \ (1 - \alpha)\mathcal{L}(\theta) + \alpha \, \mathcal{L}_{\mathrm{Aug}}(\theta) + \lambda \|\theta\|_2^2 \ . \tag{3}$$

Note that we could set $\alpha = 0$ up to a change of $(\tau_\mathtt{x}, \tau_\mathtt{y}, \mathcal{T})$. We do not adopt this presentation, however, as it would make our assumptions and results harder to state.

Our main object of interest is the generalization error of $\hat{\theta}_{\alpha,\lambda}$

$$\mathcal{G}(\alpha, \lambda) := \mathbb{E}\Big[\big\{ \varphi(X)^\top \hat{\theta}_{\alpha,\lambda} - Y \big\}^2 \Big] \ , \tag{4}$$

where (X, Y) is an independent test pair drawn from the same distribution as the training data. In contrast to previous works [LKDM24], we do not suppose that the augmentation scheme is unbiased, i.e., for some $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ it may hold that $\mathbb{E}\left[\tau_\mathtt{x}(x, y, \eta)\right] \neq x$, and $\mathbb{E}\left[\tau_\mathtt{y}(x, y, \eta)\right] \neq y$, which enables the analysis of potential bias introduced by the DA scheme. To conduct our analysis, we introduce the first moments of the DA scheme defined for any $z \in \mathbb{R}^d \times \mathbb{R}$ as

$$\mu_\mathtt{x}(z) = \mathbb{E}\left[\varphi(\tau_\mathtt{x}(z, \eta))\right] \ , \quad \mu_\mathtt{y}(z) = \mathbb{E}\left[\tau_\mathtt{y}(z, \eta)\right] \ , \tag{5}$$

as well as the corresponding covariance operators,

$$\Lambda(z) = \mathrm{Cov}[\varphi(\tau_\mathtt{x}(z, \eta))] \ , \qquad \Omega(z) = \mathrm{Cov}[\varphi(\tau_\mathtt{x}(z, \eta)), \tau_\mathtt{y}(z, \eta)] \ . \tag{6}$$

We also define their empirical averages over the training samples,

$$\Lambda(\mathtt{Z}) = \frac{1}{n} \sum_{i=1}^n \Lambda(\mathtt{Z}_i) \ , \quad \text{and} \quad \Omega(\mathtt{Z}) = \frac{1}{n} \sum_{i=1}^n \Omega(\mathtt{Z}_i) \ ,$$

as well as their population counterparts,

$$\overline{\Lambda} = \mathbb{E}\left[\Lambda(z)\right] \ , \quad \overline{\Omega} = \mathbb{E}\left[\Omega(z)\right] \ .$$

These quantities are of natural interest, as they quantify the implicit regularization effect induced by the DA scheme [Bis95, LKDM24]. Indeed, writting a bias variance decomposition on (2) gives,

$$\mathcal{L}_{\mathrm{Aug}}(\theta) = \frac{1}{n} \|\mu_\mathtt{y}(\mathtt{Z}) - \theta^\top \mu_\mathtt{x}(\mathtt{Z})\|_2^2 - 2\theta^\top \Omega(\mathtt{Z}) + \theta^\top \Lambda(\mathtt{Z})\theta + \mathrm{c} \ ,$$

where here, c is a constant that doesn't depend on θ. Thus the first order condition in (3) yields,

$$\hat{\theta}_{\alpha,\lambda} = \mathrm{R}_{\alpha,\lambda}\mathrm{H}_\alpha \ , \tag{7}$$

where

$$\mathrm{H}_\alpha = ((1-\alpha)\mathrm{H} + \alpha\mathrm{H}' + \alpha\Omega(\mathtt{Z})) \ , \qquad \mathrm{R}_{\alpha,\lambda} = ((1-\alpha)\mathrm{C} + \alpha\mathrm{C}' + \alpha\Lambda(\mathtt{Z}) + \lambda \mathrm{I}_p)^{-1} \ , \tag{8}$$

and

$$\mathrm{C} = \frac{1}{n}\varphi(\mathtt{X})\varphi(\mathtt{X})^\top \ , \quad \mathrm{H} = \frac{1}{n}\varphi(\mathtt{X})\mathtt{Y}^\top \ ,$$
$$\mathrm{C}' = \frac{1}{n}\mu_{\mathtt{x}}(\mathtt{Z})\mu_{\mathtt{x}}(\mathtt{Z})^\top \ , \quad \mathrm{H}' = \frac{1}{n}\mu_{\mathtt{x}}(\mathtt{Z})\mu_{\mathtt{y}}(\mathtt{Z})^\top \ .$$

The matrix $\mathrm{R}_{\alpha,\lambda}$ in (8) is the resolvent associated with the effective empirical covariance $(1-\alpha)\mathrm{C} + \alpha\mathrm{C}' + \alpha\Lambda(\mathtt{Z})$: it therefore governs the stability and sensitivity of $\hat{\theta}_{\alpha,\lambda}$ from (7). Resolvent matrices are a central object in Random Matrix Theory, as the Stieltjes transform of the empirical spectral distribution can be expressed in terms of normalized traces of $\mathrm{R}_{\alpha,\lambda}$, which makes it a convenient tool to characterize eigenvalue distributions in high dimension. Data augmentation modifies this resolvent in two key ways. First, it effectively enriches the dataset by incorporating $\mu_{\mathtt{x}}(\mathtt{Z})$. Second, it introduces an extra data-dependent regularization term $\Lambda(\mathtt{Z})$ which is typically anisotropic and inflates directions in feature space in proportion to the augmentation-induced variability.

Our analysis of the generalisation error is made under the proportional regime, i.e.,

H2. *We assume that $p = p(n)$ and*

$$\limsup_{n\to\infty} p/n < \infty \ .$$

H2 is standard in high-dimensional asymptotics and allows the covariate dimension p to scale at most linearly with the sample size n. In particular, it covers both underparameterized and overparameterized regimes, including cases where $p > n$. Although it is not the focus of this paper, our results also cover the classical regime $p = \mathcal{O}(1)$.

We shall work under the following assumptions on the data distribution and the DA scheme:

H3. *We assume that:*

(i) for any 1-Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, the random variable $f(\mathtt{X}_1)$ is 1-sub-Gaussian.

(ii) the feature maps $\varphi : \mathbb{R}^d \to \mathbb{R}^p$ and $\varphi_\star : \mathbb{R}^d \to \mathbb{R}^{p_\star}$ are $\mathtt{L}_\varphi$-Lipschitz continuous, with $\mathtt{L}_\varphi = \mathcal{O}(1)$ independent of n.

H3 formalizes the concentration of measure conditions supposed throughout the paper. In particular the concentration hypothesis in **H**3-(i) is commonly refered to as the *Lipschitz concentration* property, for which we refer the reader to [LC18, LC20, LC21] for more background and details. Naturally, combining **H**1 and **H**3 implies that the pairs $\mathtt{Z}_i = (\mathtt{X}_i, \mathtt{Y}_i)$ satisfies **H**3-(i) with a sub-Gaussian parameter bounded by $1 + \mathtt{L}_\varphi \|\theta_\star\| + \sigma^2$. As a simple example of distribution for which **H**3 holds, one can consider any Gaussian distribution $\mathrm{N}(0, \mathbf{A})$, where $\|\mathbf{A}\|_{\mathrm{op}} \leqslant 1$, [LC18].

We now impose a mild stability condition on the augmentation scheme, requiring its first and second-order statistics to vary smoothly with the data.

H4. *We assume that $\mu_{\mathtt{x}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^p$, $\mu_{\mathtt{y}} : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ defined in (5), and $\Lambda : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{p\times p}$, $\Omega : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{p\times 1}$ defined in (6) are all Lipschitz continuous, with Lipschitz constants universally bounded in n. In addition, we assume that for a constant $c > 0$ independant of n,*

$$\mathrm{Var}(\Lambda(\mathtt{Z}_1)) + \mathrm{Var}(\Omega(\mathtt{Z}_1)) \leqslant \mathrm{c} \ . \tag{9}$$

Note that **H**4 is a mere technical assumption that serves only the purpose of simplicity, we expect that our results can be extended under relaxed versions of this assumption as well as **H**3. Indeed, previous works in Random Matrix Theory suggest that we can extend the results developped in Proposition 1 under only finite moment assumptions on the entries of $\mathtt{Z}$ [LP11]. The bounds in (9) are necessary to ensure the stability of the data augmentation scheme with respect to perturbations of the dataset. Without such a stability condition, any general analysis of the generalization error $\mathcal{G}(\alpha, \lambda)$ appears very challenging.

Our setting covers a wide range of augmentation pipelines used in practice, including label-preserving perturbations such as additive Gaussian noise, random feature dropout/masking, and small geometric or temporal deformations, as well as more structured transformations that explicitly modify the target. In particular, we significantly enlarge the class of DA schemes that can be analyzed compared to prior works, e.g., [LKDM24] requires that $\tau_{\mathtt{y}}((x,y),\cdot) = y$ and $\tau_{\mathtt{x}}((x,y),\cdot) = \tau'_{\mathtt{x}}(x,\cdot)$ for all $(x,y) \in \mathbb{R}^d \times \mathbb{R}$ and a map $\tau'_{\mathtt{x}}$. Below, we describe a simple salt-and-pepper DA scheme in the identity feature map case, for which $\Lambda(\mathtt{Z})$ and $\Omega(\mathtt{Z})$ admit closed-form expressions, and for which **H** 4 can easily be verified. Additional examples of schemes are given in Section B:

Example: Salt-and-pepper noise injection. Let $k \in \mathbb{N}$, and fix a masking probability $\mathfrak{m} : \mathbb{R}^d \times \mathbb{R} \to [0,1]$ and a replacement map $s : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^{d\times k}$. We consider the augmentation scheme with transformations for any $z = (x,y) \in \mathbb{R}^d \times \mathbb{R}$,

$$\begin{aligned}\tau_{\mathtt{x}}(z,\eta) &= x \odot m_z + s(z)(\eta_2 - \eta_2 \odot m_z)\ ,\\ \tau_{\mathtt{y}}(z,\eta) &= y\ ,\end{aligned}$$

where $\eta = (\{U_i\}_{i=1}^d, \eta_2)$ are independent random variables such that $\{U_i\}_{i=1}^d$ is i.i.d. with distribution $\mathrm{Unif}([0,1])$, m_z is the vector with i-th component given by $\mathbb{1}\{U_i \leqslant \mathfrak{m}(z)\}$ and η_2 is a zero-mean k-dimensional random vector with $\mathbb{E}\left[\eta_2\eta_2^\top\right] = \mathrm{I}_k$. This construction yields $\Omega(\mathtt{Z}_i) = 0$ for all $i \leq n$, and

$$\Lambda(\mathtt{Z}_i) = (1 - \mathfrak{m}(\mathtt{Z}_i))\left[\mathfrak{m}(\mathtt{Z}_i)\mathrm{Diag}(\mathtt{X}_i\mathtt{X}_i^\top) + s(\mathtt{Z}_i)s(\mathtt{Z}_i)^\top\right]\ .$$

We refer to Section B for details of the computation of $\Lambda(\mathtt{Z}_i)$.

3 Main results

We start by introducing the notion of stochastic domination, introduced in [EKY13]. In our work, this definition is used to simplify the presentation of our results by retaining only the dominant term in n in our bounds. This convention is also adopted in [SCDL23, SDCL24].

Definition 1. *We say the sequence of random variables $(U_n)_{n\in\mathbb{N}}$ is stochastically dominated by the sequence of random variables $(V_n)_{n\in\mathbb{N}}$, and write $U_n \prec_s V_n$, if for all $\varepsilon, D > 0$, it holds for n large enough that*

$$\mathbb{P}\left(U_n \geqslant n^{\varepsilon} V_n\right) \leqslant n^{-D}\ .$$

To state our main result in a concise and informative way, we first introduce the following random matrices

$$\Phi_i = (\varphi(\mathtt{X}_i), \mu_{\mathtt{x}}(\mathtt{Z}_i))\ ,\ \Psi_i = (\mathtt{Y}_i, \mu_{\mathtt{y}}(\mathtt{Z}_i))\ , \tag{10}$$

which are used to define the following dilation matrices:

$$\mathrm{A}_{\alpha,\lambda} = \begin{pmatrix} \mathrm{tr}\left(\Sigma\,\mathrm{R}_{\alpha,\lambda}\right) & \mathrm{tr}\left(\Sigma'\,\mathrm{R}_{\alpha,\lambda}\right) \\ \mathrm{tr}\left(\Sigma'\,\mathrm{R}_{\alpha,\lambda}\right) & \mathrm{tr}\left(\Sigma''\,\mathrm{R}_{\alpha,\lambda}\right)\end{pmatrix}\ ,\quad \mathrm{B}_{\alpha,\lambda} = \mathrm{W}_\alpha\left(\mathrm{I}_2 + \frac{1}{n}\mathrm{W}_\alpha\mathbb{E}\left[\mathrm{A}_{\alpha,\lambda}\right]\mathrm{W}_\alpha\right)^{-1}\mathrm{W}_\alpha\ , \tag{11}$$

$$\mathrm{C}_{\alpha,\lambda} = \begin{pmatrix} \mathrm{tr}\left(\Sigma\,\mathrm{R}_{\alpha,\lambda}\Sigma\,\mathrm{R}_{\alpha,\lambda}\right) & \mathrm{tr}\left(\Sigma'\,\mathrm{R}_{\alpha,\lambda}\Sigma\,\mathrm{R}_{\alpha,\lambda}\right) \\ \mathrm{tr}\left(\Sigma'\,\mathrm{R}_{\alpha,\lambda}\Sigma\,\mathrm{R}_{\alpha,\lambda}\right) & \mathrm{tr}\left(\Sigma''\,\mathrm{R}_{\alpha,\lambda}\Sigma\,\mathrm{R}_{\alpha,\lambda}\right)\end{pmatrix}\ ,\quad \mathrm{D}_{\alpha,\lambda} = \frac{1}{n}\mathrm{B}_{\alpha,\lambda}\mathbb{E}\left[\mathrm{C}_{\alpha,\lambda}\right]\mathrm{B}_{\alpha,\lambda}\ , \tag{12}$$

where $\mathrm{W}_\alpha = \mathrm{Diag}(\sqrt{1-\alpha}, \sqrt{\alpha})$ is a diagonal weight matrix, and $\Sigma, \Sigma', \Sigma''$ are the feature covariance matrices defined as

$$\Sigma = \mathbb{E}\left[\varphi(\mathtt{X}_1)\varphi(\mathtt{X}_1)^\top\right]\ ,\quad \Sigma' = \mathbb{E}\left[\varphi(\mathtt{X}_1)\mu_{\mathtt{x}}(\mathtt{Z}_1)^\top\right]\ ,\quad \Sigma'' = \mathbb{E}\left[\mu_{\mathtt{x}}(\mathtt{Z}_1)\mu_{\mathtt{x}}(\mathtt{Z}_1)^\top\right]\ .$$

The matrices introduced in (11) and (12) extend the dilation factors that appear in previous works [Cho22] for the non-augmented setting. They arise naturally in our analysis through a fixed-point system satisfied by the resolvent matrix $\mathrm{R}_{\alpha,\lambda}$ defined in (8), together with concentration estimates for high-dimensional quadratic forms. More concretely, in the proportional regime the resolvent is asymptotically unchanged when the contribution of a fixed sample (say Φ_1) is removed from the empirical covariance, so one can compare the full resolvent to its 'leave-one-out'counterpart and enforce self-consistency between the two descriptions; typically through the use of the Woodbury matrix identity, which yields the said fixed-point system.

With these notations in place, the deterministic equivalent can be written in terms of the following population quantities:

$$\overline{\Sigma}_{\alpha,\lambda} = \mathbb{E}\left[\Phi_1 \mathrm{B}_{\alpha,\lambda} \Phi_1^\top\right] \ , \quad \overline{\Sigma}'_{\alpha,\lambda} = \mathbb{E}\left[\Phi_1 \mathrm{D}_{\alpha,\lambda} \Phi_1^\top\right] \ ,$$
$$\overline{\Gamma}_{\alpha,\lambda} = \mathbb{E}\left[\Phi_1 \mathrm{B}_{\alpha,\lambda} \Psi_1^\top\right] \ , \quad \overline{\Gamma}'_{\alpha,\lambda} = \mathbb{E}\left[\Phi_1 \mathrm{D}_{\alpha,\lambda} \Psi_1^\top\right] \ , \quad \overline{\gamma}_{\alpha,\lambda} = \mathbb{E}\left[\Psi_1 \mathrm{D}_{\alpha,\lambda} \Psi_1^\top\right] \ .$$

Note that because true and artificial data are stacked together in (10), each of these quantities can be expanded into linear combinations of population (cross-)covariance terms involving $\varphi(\mathtt{X}_1)$, $\mu_{\mathtt{x}}(\mathtt{Z}_1)$, $\mathtt{Y}_1$, and $\mu_{\mathtt{y}}(\mathtt{Z}_1)$ only, where the linear coefficient terms are given by the entries of $\mathrm{B}_{\alpha,\lambda}$ in (11), and $\mathrm{D}_{\alpha,\lambda}$ in (12) respectively. Our Theorem below gives an explicit deterministic equivalent of the generalization error introduced in (4), together with an explicit non-asymptotic control of the approximation error.

Theorem 1. *Let $\alpha \in [0,1]$ and $(\lambda_n)_{n\in\mathbb{N}}$, such that for all n, $\lambda_n \geqslant \mathrm{c}/n$ for some $\mathrm{c} > 0$. Assume that **H**1, **H**2, **H**3 and **H**4 hold, and define*

$$\overline{\mathcal{G}}(\alpha,\lambda_n) = \theta_\star^\top \Sigma_{\star\star} \theta_\star - 2\theta_\star^\top \Sigma_\star \overline{\theta}_{\alpha,\lambda_n} + \overline{\gamma}_{\alpha,\lambda} + \overline{\theta}_{\alpha,\lambda_n}^\top \{\overline{\Sigma}'_{\alpha,\lambda} + \Sigma\} \overline{\theta}_{\alpha,\lambda_n} - 2\overline{\Gamma}'^\top_{\alpha,\lambda_n} \overline{\theta}_{\alpha,\lambda_n} + \sigma^2 \ ,$$

where

$$\Sigma_\star = \mathbb{E}\left[\varphi_\star(\mathtt{X}_1)\varphi(\mathtt{X}_1)^\top\right] \ , \quad \Sigma_{\star\star} = \mathbb{E}\left[\varphi_\star(\mathtt{X}_1)\varphi_\star(\mathtt{X}_1)^\top\right] \ .$$
$$\textit{and} \quad \overline{\theta}_{\alpha,\lambda_n} = \left(\overline{\Sigma}_{\alpha,\lambda} + \alpha\overline{\Lambda} + \lambda_n \mathrm{I}_d\right)^{-1} \left(\overline{\Gamma}_{\alpha,\lambda} + \alpha\overline{\Omega}\right) \ . \tag{13}$$

Then the generalization error $\mathcal{G}$ defined in (4) satisfies

$$\left|\mathcal{G}(\alpha,\lambda_n) - \overline{\mathcal{G}}(\alpha,\lambda_n)\right| \prec_s \frac{1}{\lambda_n^{9/2}\sqrt{n}} + \frac{\|\theta_\star\|_2}{\lambda_n^{7/2}\sqrt{n}} \ .$$

The proof of Theorem 1 is defered to Section F.

The previous result provides a non-asymptotic control with an explicit dependence on the ridge parameter λ_n. In particular, the remainder term scales as $\lambda_n^{-9/2}$. This explicit λ_n-dependence is important in the proportional regime, where λ_n may be small. In particular, our results may be extended to the Ordinary Least-Squares regression, by allowing λ_n to converge to 0 slowly enough. Note that this dependence in $\lambda_n^{-9/2}$ is competitive with the sharpest available bounds in the literature [Cho22, BPH23, MM22a].

To gain some qualitative intuition, we specialize the theorem to the well-specified case $\varphi = \varphi_\star$ together with an unbiased, label-preserving augmentation scheme, so that $\mu_{\mathtt{x}}(\mathtt{Z}_1) = \varphi(\mathtt{X}_1)$ and $\mu_{\mathtt{y}}(\mathtt{Z}_1) = \mathtt{Y}_1$. In that regime, the mean augmented features coincide with the original ones, hence $\mathrm{C}' = \mathrm{C}$ and $\mathrm{H}' = \mathrm{H}$, while the feature covariances satisfy $\Sigma = \Sigma' = \Sigma'' = \Sigma_\star = \Sigma_{\star\star}$. We also introduce

$$\beta_{\alpha,\lambda_n} = \mathbf{1}^\top \mathrm{B}_{\alpha,\lambda_n} \mathbf{1} \ , \quad \delta_{\alpha,\lambda_n} = \mathbf{1}^\top \mathrm{D}_{\alpha,\lambda_n} \mathbf{1} \ .$$

In this regime, the equivalent of the regressor $\overline{\theta}_{\alpha,\lambda_n}$ in (13) reduces to

$$\overline{\theta}_{\alpha,\lambda_n} = \left(\beta_{\alpha,\lambda_n}\Sigma + \alpha\overline{\Lambda} + \lambda_n \mathrm{I}_d\right)^{-1} \left(\beta_{\alpha,\lambda_n}\Gamma + \alpha\overline{\Omega}\right) \ , \quad \text{where} \quad \Gamma = \mathbb{E}\left[\varphi(\mathtt{X}_1)\mathtt{Y}_1\right] \ .$$

Moreover, the equivalent of the generalization error simplifies to

$$\overline{\mathcal{G}}(\alpha,\lambda_n) = (1+\delta_{\alpha,\lambda_n}) \left[\left(\overline{\theta}_{\alpha,\lambda_n} - \theta_\star\right)^\top \Sigma \left(\overline{\theta}_{\alpha,\lambda_n} - \theta_\star\right) + \sigma^2\right] \ .$$

This expression makes the role of DA particularly transparent. In this simplified setting, DA acts as an additional anisotropic regularization through $\alpha\overline{\Lambda}$, on top of the usual isotropic ridge shift $\lambda_n \mathrm{I}_d$. Hence, DA can improve over standard ridge when $\overline{\Lambda}$ is large along directions that carry little or no predictive signal, while remaining small along directions aligned with $\theta_\star$. Intuitively, the augmentation should inject variability in nuisance directions rather than in informative ones (e.g. directions associated with symetries in the distribution). Conversely, if $\overline{\Lambda}$ penalizes signal-carrying directions too strongly, the resulting bias increase can dominate the variance reduction and worsen the generalization error.

In what follows, we present intermediate results that we derive to complete the proof of Theorem 1 and that we believe are of independent interest. In particular, we provide in our next result a deterministic equivalent for $\hat{\theta}_{\alpha,\lambda_n}$ and $\hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n}$. Indeed, expending (4), we get

$$\mathcal{G}(\alpha, \lambda_n) = \hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n} - 2\theta_\star^\top \Sigma_\star \hat{\theta}_{\alpha,\lambda_n} + \theta_\star^\top \Sigma_{\star\star} \theta_\star + \sigma^2 \ , \tag{14}$$

The proof of Theorem 1 mostly relies on deterministic equivalents for the first two terms in the decomposition above. This is the purpose of Proposition 1. There, we show that these terms are approximated by deterministic quantities involving the population covariances of both the original and augmented data.

We first define the deterministic equivalent of the resolvent

$$\overline{\mathrm{R}}_{\alpha,\lambda_n} = \left(\overline{\Sigma}_{\alpha,\lambda_n} + \alpha\overline{\Lambda} + \lambda_n \mathrm{I}_d\right)^{-1} \ ,$$

as well as the following notation for the quadratic form in (14) and its deterministic equivalent:

$$\begin{aligned}
\hat{\chi}_{\alpha,\lambda_n} &= \hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n} \ , \\
\overline{\chi}_{\alpha,\lambda_n} &= \overline{\theta}_{\alpha,\lambda_n}^\top \{\overline{\Sigma}'_{\alpha,\lambda_n} + \Sigma\} \overline{\theta}_{\alpha,\lambda_n} - 2\overline{\theta}_{\alpha,\lambda_n}^\top \overline{\Gamma}'_{\alpha,\lambda_n} + \overline{\gamma}_{\alpha,\lambda_n} \ .
\end{aligned}$$

Then, it holds:

Proposition 1. *Let $(\lambda_n)_{n\in\mathbb{N}}$, such that for all n, $\lambda_n \geqslant \mathrm{c}/n$, for $\mathrm{c} > 0$. Assume that **H**1, **H**2, **H**3 and **H**4 hold. Then for any $\alpha \in [0, 1]$, we have:*

1. *For any $\mathbf{A} \in \mathbb{R}^{p\times p}$*
$$\left|\mathrm{tr}\left(\mathbf{A}\left\{\mathrm{R}_{\alpha,\lambda_n} - \overline{\mathrm{R}}_{\alpha,\lambda_n}\right\}\right)\right| \prec_s \frac{\|\mathbf{A}\|_{\mathrm{op}}}{\lambda_n^{7/2}\sqrt{n}} \ .$$
2. *For any $\mathbf{a} \in \mathbb{R}^p$*
$$\left|\mathbf{a}^\top\{\hat{\theta}_{\alpha,\lambda_n} - \overline{\theta}_{\alpha,\lambda_n}\}\right| \prec_s \frac{\|\mathbf{a}\|_2}{\lambda_n^4\sqrt{n}} \ ,$$
where $\overline{\theta}_{\alpha,\lambda_n}$ was defined in (13).
3. *Finally*
$$\left|\hat{\chi}_{\alpha,\lambda_n} - \overline{\chi}_{\alpha,\lambda_n}\right| \prec_s \frac{1}{\lambda_n^{9/2}\sqrt{n}} \ .$$

The previous result provides a deterministic equivalent for resolvents of sample covariances matrices with non-independent samples, as for all i, the pair $(\varphi(\mathtt{X}_i), \mu_{\mathtt{x}}(\mathtt{Z}_i))$ is allowed to have an arbitrary joint distribution. To our knowledge, such relaxation of the independance assumption has only been considered in [MHD25], also in the context of data-augmentation.

4 Simulations

In this section, we use synthetic datasets to assess the finite-sample accuracy of Theorem 1 and Proposition 1. We find that the proposed deterministic equivalents closely match their empirical counterparts across a broad range of dimensions and sample sizes, provided the regularization level λ_n is taken sufficiently large, which is consistent with the requirements suggested by our error bounds. More unexpectedly, and contrary to common intuition, our simulations show that DA impacts the model's bias in a non-monotonic way, under feature misspecification. In particular, when the augmentation scheme substantially alters the covariate structure of the distribution, DA can avoid (or even mitigate) the bias increase one might anticipate. This contrasts with prior results in the correctly specified setting, where stronger regularization is shown to monotonically increase bias [BPH23, MM22a, LKDM24]. Additional experiments on real data, including a more practical inpainting task on MNIST, are deferred to Section A.

In all experiments, we consider a synthetic regression problem in which the training data $(\mathtt{X}_i, \mathtt{Y}_i)_{i=1}^n$ are generated as follows. The covariates are i.i.d. Gaussian, $\mathtt{X}_i \sim \mathcal{N}(0, \Sigma_{\mathtt{X}})$, and the responses are generated according to **H**1. To mimic the strong anisotropy often observed in high-dimensional

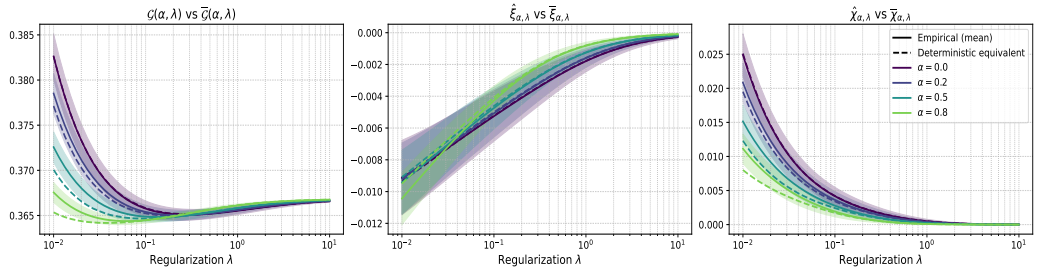


Figure 1: **Deterministic equivalents accuracy for various λ and α.** Comparison between empirical estimates (solid lines) and deterministic equivalents (dashed lines) as the regularization λ varies, for multiple augmentation strengths α (colors). Shaded bands represent ± 1 empirical standard deviation over our Monte-Carlo approximations of $\mathbb{E}\left[\mathcal{G}(\alpha,\lambda)\right]$, $\mathbb{E}\left[\xi_{\alpha,\lambda}\right]$ and $\mathbb{E}\left[\chi_{\alpha,\lambda}\right]$.

learning, we take $\Sigma_{\mathtt{X}} = Q\,\mathrm{Diag}\big((k^{-1})_{k=1}^{1000}\big)\,Q^\top$, i.e., a power-law spectrum, where Q is Haar-distributed on the orthogonal group of $\mathbb{R}^{1000}$.

To highlight the effects of feature misspecification, we use $\varphi_\star : \mathbb{R}^{1000} \to \mathbb{R}^{700}$ to generate the labels, and $\varphi : \mathbb{R}^{1000} \to \mathbb{R}^{700}$ to predict the labels, as randomly initialized MLPs, of respective hidden layer sizes $[1000, 1000, 1000]$ and $[500, 500]$. Note that this choice of feature map is made to ensure strong misspecification. We set the ground-truth parameter to $\theta_\star = (1, \dots, 1)^\top / \sqrt{700} \in \mathbb{R}^{700}$ and add Gaussian label noise $\varepsilon \sim \mathcal{N}(0, 0.5)$.

We consider a *salt-and-pepper* augmentation scheme, defined by

$$\tau_{\mathtt{X}}(x, y) \;=\; \mathfrak{m} \odot x \;+\; (1 - \mathfrak{m}) \odot \eta, \qquad \tau_{\mathtt{Y}}(x, y) \;=\; y \;,$$

where $\mathfrak{m} \in \{0, 1\}^{1000}$ is a random mask of Bernoulli random variables with parameter 0.2, and η is an independent isotropic noise vector of variance 0.5. Although simple, this transformation substantially alters the features, and yields non-trivial functions $\mu_{\mathtt{X}}$ and Λ. These quantities do not admit a simple closed form for nonlinear φ and are therefore computed via Monte–Carlo estimation in our experiments.

Our setup also allows us to disentangle how DA affects the bias and variance contributions to generalization. Indeed, the generalization error can be decomposed as

$$\mathbb{E}\left[\mathcal{G}(\alpha, \lambda_n)\right] = \mathrm{bias}^2(\alpha, \lambda_n) + \mathrm{V}(\alpha, \lambda_n) \;, \tag{15}$$

where

$$\mathrm{bias}^2(\alpha, \lambda_n) = \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}_{\alpha,\lambda_n}\right]^\top \varphi(X) - \theta_\star^\top \varphi_\star(X)\right)^2\right] \;,$$

which rewrites as

$$\mathrm{bias}^2(\alpha, \lambda_n) = \mathbb{E}[\hat{\theta}_{\alpha,\lambda_n}]^\top \Sigma \mathbb{E}[\hat{\theta}_{\alpha,\lambda_n}] + \theta_\star^\top \Sigma_{\star\star} \theta_\star - 2\theta_\star^\top \Sigma_\star \mathbb{E}[\hat{\theta}_{\alpha,\lambda_n}] \;.$$

In light of Item 2 of Proposition 1, we thus introduce

$$\overline{\mathrm{bias}}^2(\alpha, \lambda_n) = \overline{\theta}_{\alpha,\lambda_n}^\top \Sigma \overline{\theta}_{\alpha,\lambda_n} + \theta_\star^\top \Sigma_{\star\star} \theta_\star - 2\theta_\star^\top \Sigma_\star \overline{\theta}_{\alpha,\lambda_n} \;,$$

ensuring that for a large enough constant $\mathrm{c} > 0$

$$\left|\mathrm{bias}^2(\alpha, \lambda_n) - \overline{\mathrm{bias}}^2(\alpha, \lambda_n)\right| \leqslant \frac{\mathrm{c}}{\lambda_n^{7/2}\sqrt{n}} \;,$$

Note that the variance term $\mathrm{V}(\alpha, \lambda_n)$ does not admit a simple closed-form expression, still it is estimated as the remainder in (15), where both the generalization error and the bias are estimated thanks to Monte-Carlo approximations.

Figure 1 reports the accuracy of our deterministic equivalents for the generalization error $\mathcal{G}(\alpha, \lambda_n)$ and for the auxiliary quantities $\xi_{\alpha,\lambda_n} = \theta_\star^\top \Sigma_\star \hat{\theta}_{\alpha,\lambda_n}$ and χ_{α,λ_n} introduced in Theorem 1 and Proposition 1. Overall, the deterministic predictions closely track their empirical counterparts across the range of regularization parameters considered. Consistent with the error bounds, the approximation

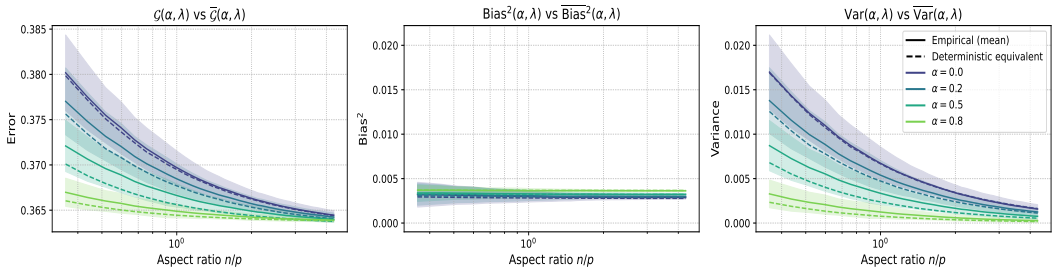


Figure 2: **Bias–variance decomposition for various aspect ratios.** Empirical (solid) and deterministic-equivalent (dashed) learning curves as the aspect ratio n/p varies (log scale), at a fixed regularization level $\lambda = 0.05$. Shaded bands show ± 1 empirical standard deviation over our Monte-Carlo approximations of $\mathcal{G}(\alpha, 0.05)$, $\mathrm{bias}^2(\alpha, 0.05)$ and $\mathrm{V}(\alpha, 0.05)$.

deteriorates as $\lambda_n \to 0$. For this reason, we restrict attention to $\lambda_n \geq 10^{-2}$, where the match is uniformly tight. In this regime, the proposed salt-and-pepper augmentation yields a clear improvement over the non-augmented baseline across all $\lambda_n \in [10^{-2}, 1]$.

Figure 2 illustrates how augmentation impacts the bias and variance components of the prediction error, across multiple choices of α and the aspect ratio p/n, at the fixed regularization level $\lambda_n = 0.05$. As expected, increasing the diversity of samples through DA reduces the variance term, consistent with the view of data augmentation as an implicit regularizer. More surprisingly, we do not observe a systematic increase of the bias, even though the augmented covariates can differ substantially from the original ones. Across a broad range of practitioner-relevant augmentation strengths, the bias remains largely unchanged, and the overall gain is driven primarily by variance reduction. A marked bias increase appears only when the augmentation is pushed to unrealistically strong regimes.

This contrasts with previous observations in correctly specified settings, where stronger regularization effects are typically accompanied by increased bias, and only appears in misspecified feature models: since the estimator is already biased due to mismatch between φ and $\varphi_\star$, the additional bias induced by augmentation can be negligible compared to this intrinsic misspecification bias. These results indicate that DA may exhibit little to no bias–variance trade-off, acting predominantly as a variance-reduction mechanism.

5 Conclusion

We developed a high-dimensional theory for supervised feature regression trained with arbitrary data augmentation, allowing for substantial perturbations of the covariates. In the proportional regime, we derived deterministic equivalents for the augmented ridge estimator and its out-of-sample error, together with explicit non-asymptotic error controls in terms of n and the ridge parameter λ. Technically, our analysis relies on resolvent methods that accommodate the non-i.i.d. structure induced by augmentation via the coupled objects $(\varphi(\mathtt{X}_i), \mu_{\mathtt{x}}(\mathtt{Z}_i))$.

Beyond prediction formulas, our results provide a mechanistic view of how augmentation changes learning under feature misspecification: it alters the effective covariance through a mean-feature component and an additional, generally anisotropic variance term. Our simulations validate the accuracy of the deterministic equivalents and suggest that, in misspecified feature models, augmentation can improve generalization over a wide range of λ while not necessarily increasing the bias, so that gains are driven primarily by variance reduction.

Yet, our results cannot be directly applied by practitioners to predict when a DA scheme is beneficial in the finite-sample setting. A key open direction is to turn our population-level formulas into a practical, data-driven and quantitative criterion that could be used by practitioners.

References

[Ada15] Radoslaw Adamczak. A note on the hanson–wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20:1–13, 2015.

[BHMM19] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32):15849–15854, 2019.

[Bis95] Christopher M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

[BKR22] Markus Bayer, Marc-André Kaufhold, and Christian Reuter. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39, 2022.

[BLLT20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[BPH23] David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 4132–4179. PMLR, 12–15 Jul 2023.

[CDL20] Shuxiao Chen, Edgar Dobriban, and Jane H. Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

[Cho22] Clément Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv:2211.13044*, 2022.

[CZM+19] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

[DL20] Oussama Dhifallah and Yue M. Lu. A precise performance analysis of learning with random features. *arXiv preprint*, 2020.

[EKY13] László Erdős, Antti Knowles, and Horng-Tzer Yau. Averaging fluctuations in resolvents of random band matrices. *Annales Henri Poincaré*, 14(8):1837–1926, 2013.

[FGW+21] Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

[FSTF16] Alhussein Fawzi, Horst Samulowitz, Deepak Turaga, and Pascal Frossard. Adaptive data augmentation for image classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3688–3692. Ieee, 2016.

[GLR+22] Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. The gaussian equivalence of generative models for learning with shallow neural networks. In Joan Bruna, Jan Hesthaven, and Lenka Zdeborova, editors, *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pages 426–471. PMLR, 2022.

[GMRM23] Florentin Guth, Brice Ménard, Gaspar Rochette, and Stéphane Mallat. A rainbow in deep network black boxes. *arXiv preprint arXiv:2305.18512*, 2023.

[HL23] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.

[HS81] H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Review*, 23(1):53–60, 1981.

[HZYN20] Ryuichiro Hataya, Jan Zdenek, Kazuki Yoshizoe, and Hideki Nakayama. Faster autoaugment: Learning augmentation strategies using backpropagation. In *European conference on computer vision*, pages 1–16. Springer, 2020.

[ITGP+23] Guillermo Iglesias, Edgar Talavera, Ángel González-Prieto, Alberto Mozo, and Sandra Gómez-Canaval. Data augmentation techniques in time series domain: a survey and taxonomy. *Neural Computing and Applications*, 35(14):10123–10145, 2023.

[IU21] Brian Kenji Iwana and Seiichi Uchida. An empirical survey of data augmentation for time series classification with neural networks. *Plos one*, 16(7):e0254841, 2021.

[KBMB24] Teerath Kumar, Rob Brennan, Alessandra Mileo, and Malika Bendechache. Image data augmentation approaches: A comprehensive survey and future directions. *IEEE Access*, 2024.

[LC18] Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices, 2018.

[LC20] Cosme Louart and Romain Couillet. Concentration of solutions to random equations with concentration of measure hypotheses, 2020.

[LC21] Cosme Louart and Romain Couillet. Spectral properties of sample covariance matrices arising from random matrices with independent non identically distributed columns, 2021.

[LKDM24] Chi-Heng Lin, Chiraag Kaushik, Eva L. Dyer, and Vidya Muthukumar. The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85, 2024.

[LKK+19] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *Advances in neural information processing systems*, 32, 2019.

[LP11] Olivier Ledoit and Sandrine Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151:233–264, 2011.

[MHD25] Lucas Morisset, Adrien Hardy, and Alain Durmus. Non-asymptotic analysis of data augmentation for precision matrix estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Spotlight Presentation.

[MM22a] Song Mei and Andrea Montanari. 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.

[MM22b] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.

[SCDL23] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30285–30320. PMLR, 23–29 Jul 2023.

[SDCL24] Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (random) features. 2024. arXiv:2402.13999 [stat.ML], v2 (10 Jun 2024).

[SK19] Connor Shorten and Taghi M. Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):60, 2019.

[SKF21] Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101, 2021.

[VAIB+25] Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint-embedding vs reconstruction: Provable benefits of latent space prediction for self-supervised learning. In *Advances in Neural Information Processing Systems*, 2025.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

[WSY+20] Qingsong Wen, Liang Sun, Fan Yang, Xiaomin Song, Jingkun Gao, Xue Wang, and Huan Xu. Time series data augmentation for deep learning: A survey. *arXiv preprint arXiv:2002.12478*, 2020.

[WWL+25] Zaitian Wang, Pengfei Wang, Kunpeng Liu, Pengyang Wang, Yanjie Fu, Chang-Tien Lu, Charu C Aggarwal, Jian Pei, and Yuanchun Zhou. A comprehensive survey on data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 2025.

[ZVP23] Jacob Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured gaussian feature models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 42866–42897. Curran Associates, Inc., 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main scope and contributions, namely deterministic equivalents for augmented random-feature regression under possibly dependent augmented samples, together with synthetic validation; see the abstract, introduction, main results, and simulations sections.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations and scope restrictions, including the proportional-regime and concentration assumptions in the settings section, as well as the fact that the population-level formulas are not yet directly usable as a practical finite-sample criterion in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The assumptions are stated in the settings section and referenced in the main results, while the proofs and auxiliary results are provided in Appendices A, B, C, D, E, and F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiments are fully synthetic and the paper specifies the data model, feature maps, augmentation scheme, dimensions, parameter ranges, and reported quantities needed to reproduce the main numerical results; see the simulations section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use synthetic data, and the submission is accompanied by anonymized code and reproduction instructions for the reported numerical results.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section specifies the synthetic data distribution, feature maps, augmentation scheme, ambient and feature dimensions, label noise level, and the parameter ranges used in the reported plots; see the simulations section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental figures in the simulations section report shaded bands corresponding to ± 1 empirical standard deviation, and the captions explicitly state what source of variability these bands represent.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The current submission does not report the hardware, runtime, or memory requirements used for the synthetic experiments. The main reason for this being the small scale of the numerical experiments, which are reproducible on most hardwares.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work is theoretical and uses only synthetic experiments, with no human subjects, scraped data, or high-risk model release; to the best of our knowledge, it conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper is primarily foundational theory with synthetic experiments, and the current submission does not include a dedicated discussion of broader positive or negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or datasets, and the experiments use synthetic data only.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

Justification: The paper does not rely on external datasets, external code assets, or pre-trained models as experimental inputs; it uses synthetic data and cites prior literature for scientific context.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The current submission does not release new datasets, code packages, or models as assets accompanying the paper.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not an important, original, or non-standard component of the core methodology studied in this research.

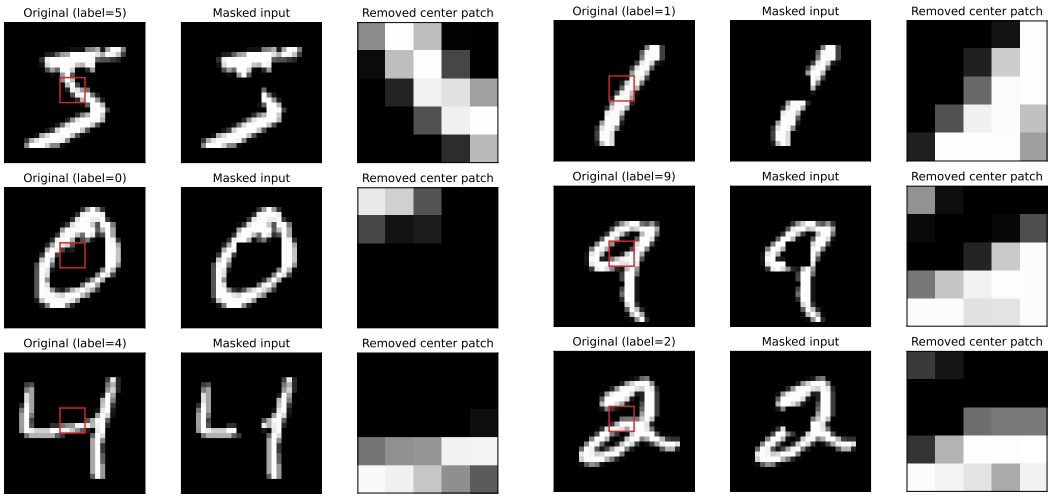
Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Additional Simulation: Inpainting on MNIST

In this section, we assess the finite-sample accuracy of Theorem 1 and Proposition 1 on an inpainting task built from the MNIST dataset. Overall, we find that the proposed deterministic equivalents closely track their empirical counterparts across a broad range of aspect ratios, not only at the level of the test risk, but also for its bias–variance decomposition.

We use the standard MNIST split, composed of $60{,}000$ training images and $10{,}000$ test images, each of size 28×28. All pixel intensities are rescaled to $[0, 1]$. For each image $I_i \in \mathbb{R}^{28\times 28}$, we remove a centered square patch of size 5×5 and aim at reconstructing the missing pixels from the remaining ones. This yields a multivariate regression problem in which the covariates $X_i \in \mathbb{R}^{759}$ are the visible pixels and the responses $Y_i \in \mathbb{R}^{25}$ are the masked central pixels. Representative masked inputs are shown in Figure 3.



(a) First three examples. (b) Last three examples.

Figure 3: **Masked MNIST examples used in the inpainting task.** Each panel displays original digits, their masked inputs, and the removed center patches. The original tall figure has been split into two halves so that three examples appear in each panel.

Although our theoretical developments were stated for scalar responses, the extension to the present vector-valued setting is immediate. Indeed, for any feature map $\varphi : \mathbb{R}^{759} \to \mathbb{R}^p$, the multivariate ridge estimator is obtained coordinatewise:

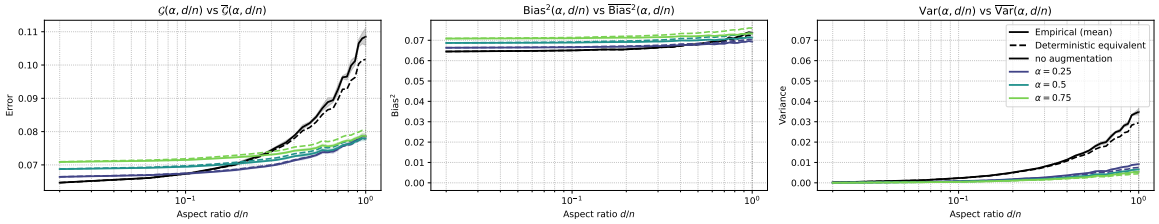
$$\widehat{\Theta}_{\alpha,\lambda_n} \in \underset{\Theta\in\mathbb{R}^{p\times 25}}{\arg\min} \frac{1}{n}\sum_{i=1}^{n}\big\|Y_i - \Theta^\top\varphi(X_i)\big\|_2^2 + \lambda_n\|\Theta\|_{\mathrm{F}}^2 = \big[\widehat{\theta}^{(1)}_{\alpha,\lambda_n},\dots,\widehat{\theta}^{(25)}_{\alpha,\lambda_n}\big],$$

where, for each coordinate $j \in \{1,\dots,25\}$, $\widehat{\theta}^{(j)}_{\alpha,\lambda_n} \in \mathbb{R}^p$ is the scalar ridge estimator obtained by regressing the j-th missing pixel on $\varphi(X_i)$. Accordingly, the deterministic equivalents derived in Theorem 1 and Proposition 1 apply coordinatewise, and the reported prediction error is the output-averaged test risk

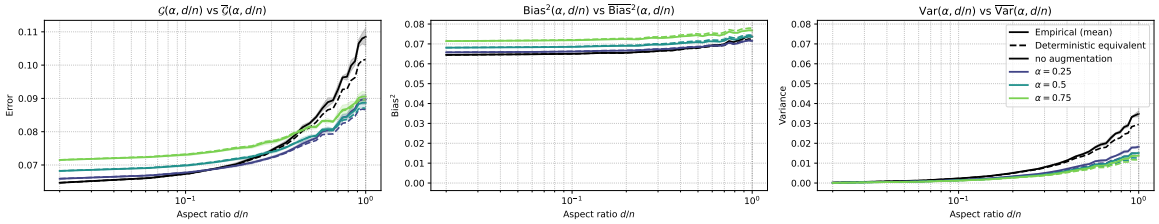
$$\mathcal{G}(\alpha,\lambda_n) := \frac{1}{25}\,\mathbb{E}\Big[\big\|Y - \widehat{\Theta}^\top_{\alpha,\lambda_n}\varphi(X)\big\|_2^2\Big] = \frac{1}{25}\sum_{j=1}^{25}\mathbb{E}\left[\Big(Y_j - \widehat{\theta}^{(j)\top}_{\alpha,\lambda_n}\varphi(X)\Big)^2\right].$$

We consider three label-preserving augmentation schemes, written in the notation of the previous sections as transformations $(\tau_{\mathtt{x}}, \tau_{\mathtt{y}})$ with $\tau_{\mathtt{y}}(z,\eta) = y$ for $z = (x, y) \in \mathbb{R}^{759} \times \mathbb{R}^{25}$:

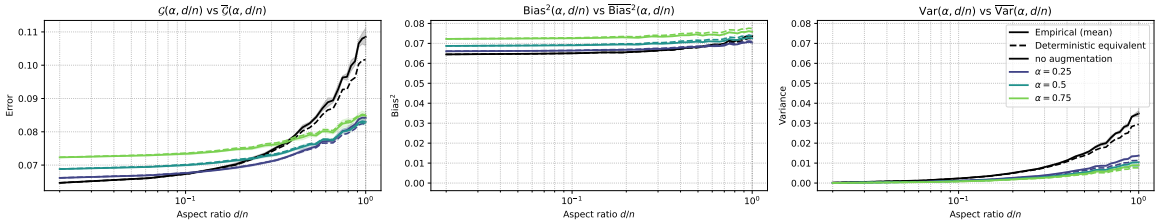
1. *Noise injection:* $\tau_{\mathtt{x}}(z,\eta) = x + \sigma_{\mathrm{aug}}\eta$, where $\eta \sim \mathbf{N}(0, \mathrm{I}_{759})$.
2. *Random masking:* $\tau_{\mathtt{x}}(z,\eta) = x \odot m$, where the entries of m are i.i.d. Bernoulli variables with keep probability parameter $q = 0.85$.
3. *Salt-and-pepper noise:* $\tau_{\mathtt{x}}(z,\eta) = x\odot m + (1-m)\odot\xi$, where again m has i.i.d. Bernoulli(q) entries with $q = 0.85$, and $\xi \sim \mathbf{N}(0, \sigma^2_{\mathrm{aug}}\mathrm{I}_{759})$.



(a) Noise injection.



(b) Random masking.



(c) Salt-and-pepper noise.

Figure 4: **MNIST inpainting with the identity feature map.** Empirical estimates and deterministic equivalents for the test error, bias, and variance as functions of the aspect ratio, for the three augmentation schemes considered in the main text and several augmentation strengths α.

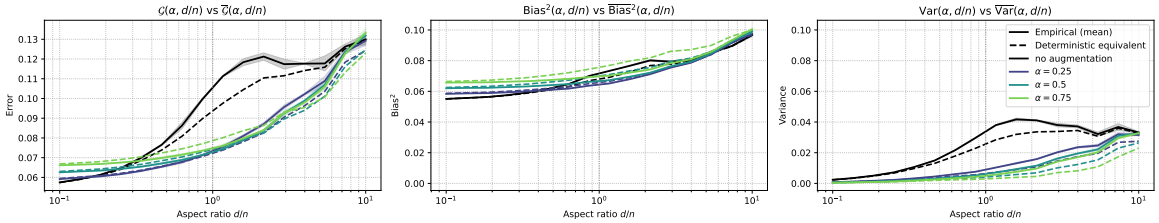
In all experiments, we set $\sigma_{\text{aug}} = 0.25$, $\lambda_n = 10^{-3}$, and $\alpha \in \{0, 0.25, 0.5, 0.75\}$, and choose $\varphi = \text{id}_{759}$, so that $p = 759$. We additionally report the same diagnostics in a second setting with a nontrivial feature map; the corresponding plots are labeled "with feature map" below. We vary the aspect ratio p/n by subsampling the training set. The population quantities entering the deterministic equivalents are estimated once from the full MNIST training set, before subsampling. We also estimate the bias–variance decomposition of the test risk at fixed aspect ratio. This allows us to disentangle how data augmentation affects the bias and variance contributions to generalization.

More precisely, for each output coordinate $j \in \{1, \dots, 25\}$, we apply the theory of Theorem 1 and Proposition 1 to the each coordinate in Y_i, i.e. the regressions of $Y_{i,j}$ on $\varphi(X_i)$, and denote the corresponding quantities by $\mathcal{G}^{(j)}(\alpha, \lambda_n)$, $\widehat{\theta}^{(j)}_{\alpha,\lambda_n}$ and $\overline{\theta}^{(j)}_{\alpha,\lambda_n}$. Then, $\mathbb{E}[\mathcal{G}^{(j)}(\alpha, \lambda_n)] = \text{bias}^2_j(\alpha, \lambda_n) + \text{V}_j(\alpha, \lambda_n)$, where

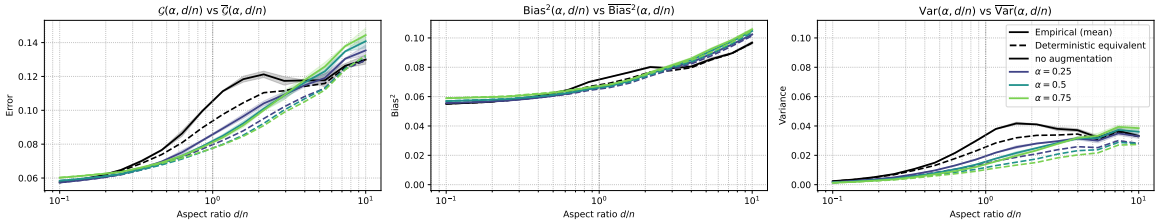
$$\begin{aligned}\text{bias}^2_j(\alpha, \lambda_n) &= \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}^{(j)}_{\alpha,\lambda_n}\right]^\top \varphi(X) - \theta^{(j)\top}_\star \varphi_\star(X)\right)^2\right] \\ &= \mathbb{E}\left[\widehat{\theta}^{(j)}_{\alpha,\lambda_n}\right]^\top \Sigma \mathbb{E}\left[\widehat{\theta}^{(j)}_{\alpha,\lambda_n}\right] + \theta^{(j)\top}_\star \Sigma^{(j)}_{\star\star} \theta^{(j)}_\star - 2\theta^{(j)\top}_\star \Sigma^{(j)}_\star \mathbb{E}\left[\widehat{\theta}^{(j)}_{\alpha,\lambda_n}\right] ,\end{aligned}$$

and its deterministic equivalent is

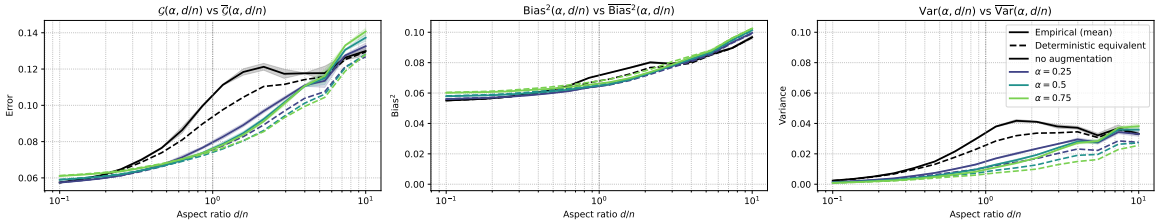
$$\overline{\text{bias}}^2_j(\alpha, \lambda_n) = \overline{\theta}^{(j)\top}_{\alpha,\lambda_n} \Sigma \overline{\theta}^{(j)}_{\alpha,\lambda_n} + \theta^{(j)\top}_\star \Sigma^{(j)}_{\star\star} \theta^{(j)}_\star - 2\theta^{(j)\top}_\star \Sigma^{(j)}_\star \overline{\theta}^{(j)}_{\alpha,\lambda_n} .$$



(a) Noise injection with feature map.



(b) Random masking with feature map.



(c) Salt-and-pepper noise with feature map.

Figure 5: **MNIST inpainting with feature map.** Same comparison as in Figure 4, after replacing the raw-pixel representation by a feature map.

We define $\mathrm{V}_j(\alpha, \lambda_n)$ and its deterministic equivalent as the remainders between the generalization error and the above bias quantities, and additionally report $\mathrm{bias}^2(\alpha, \lambda_n) = (1/25)\sum_{j=1}^{25} \mathrm{bias}_j^2(\alpha, \lambda_n)$ and $\mathrm{V}(\alpha, \lambda_n) = (1/25)\sum_{j=1}^{25} \mathrm{V}_j(\alpha, \lambda_n)$. The aspect-ratio curves for the identity feature map are reported in Figure 4. They show that the deterministic equivalents remain close to the empirical error, bias, and variance curves, and that the main effect of augmentation is a reduction of the variance term.

The corresponding experiments with feature map are displayed in Figure 5. The same qualitative behavior persists in this second setting, with close agreement between empirical quantities and deterministic equivalents across the three augmentation schemes.

We finally note that the quality of the fit is excellent in the identity feature-map setting, where the deterministic equivalents are nearly indistinguishable from their empirical counterparts across the range of aspect ratios considered. In the "with feature map" setting, the fit remains qualitatively accurate, although the discrepancy becomes more visible around the peak of the error curves. We notice on this practical setting that all DA schemes improve over the non-augmented Ridge baseline when the aspect ratio is large, i.e. in a regime where data is scarce, with the gain driven mostly by variance reduction and only a limited increase in bias, that seems not to depend on the aspect ratio.

B Sufficient conditions for H4

This section explains how **H**4 can be verified in practice by reducing it to checkable sufficient conditions. We start with a generic proposition showing how Lipschitz concentration of the augmentation randomness yields bounds on the empirical second-order quantities associated with the scheme. We then complement this result with explicit augmentation procedures for which the assumption can be checked directly.

Proposition 2. *For every $n \in \mathbb{N}$, let $\varphi : \mathbb{R}^{d(n)} \to \mathbb{R}^{p(n)}$, $\tau_{\mathtt{x}} : \mathbb{R}^{d(n)} \times \mathbb{R} \times \mathsf{E} \to \mathbb{R}^{d(n)}$ and $\tau_{\mathtt{y}} : \mathbb{R}^{d(n)} \times \mathbb{R} \times \mathsf{E} \to \mathbb{R}$. Assume that the augmentation variable η is 1-Lipschitz concentrated, and that there exist constants $\mathtt{L}_{\mathtt{x}}, \mathtt{L}_{\mathtt{y}} > 0$, independent of n, such that for all $z \in \mathbb{R}^{d(n)} \times \mathbb{R}$ and all $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathsf{E}$*

$$\|\varphi(\tau_{\mathtt{x}}(z, \boldsymbol{\eta})) - \varphi(\tau_{\mathtt{x}}(z, \boldsymbol{\eta}'))\|_2 \leqslant \mathtt{L}_{\mathtt{x}} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_2 \,,$$
$$|\tau_{\mathtt{y}}(z, \boldsymbol{\eta}) - \tau_{\mathtt{y}}(z, \boldsymbol{\eta}')| \leqslant \mathtt{L}_{\mathtt{y}} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_2 \,.$$

Then

$$\mathrm{Var}(\Lambda(\mathtt{Z})) \leqslant \mathrm{c} \frac{p}{n} \mathtt{L}_{\mathtt{x}}^4 \,, \quad \mathrm{Var}(\Omega(\mathtt{Z})) \leqslant \mathrm{c} \frac{1}{n} \mathtt{L}_{\mathtt{x}}^2 \mathtt{L}_{\mathtt{y}}^2 \,.$$

Proof. For all $z \in \mathbb{R}^{d(n)} \times \mathbb{R}$ we write

$$X_z = \varphi(\tau_{\mathtt{x}}(z, \eta)), \qquad Y_z = \tau_{\mathtt{y}}(z, \eta).$$

The Lipschitz assumptions above are with respect to the deterministic augmentation input $\boldsymbol{\eta}$. Combining them with the Lipschitz concentration of the random variable η, conditionally on any value of z, every projection $\mathbf{u}^\top X_z$, with $\|\mathbf{u}\|_2 = 1$, is sub-Gaussian with parameter of order $\mathtt{L}_{\mathtt{x}}$. Similarly, Y_z is sub-Gaussian with parameter of order $\mathtt{L}_{\mathtt{y}}$. Therefore, using the standard variance bound for sub-Gaussian random variables, for every unit vector $\mathbf{u}$, we get

$$\mathrm{Var}(\mathbf{u}^\top X_{\mathtt{Z}_1} \mid \mathtt{Z}_1) \leqslant \mathrm{c} \mathtt{L}_{\mathtt{x}}^2, \tag{16}$$
$$\mathrm{Var}(Y_{\mathtt{Z}_1} \mid \mathtt{Z}_1) \leqslant \mathrm{c} \mathtt{L}_{\mathtt{y}}^2, \tag{17}$$

almost surely. We first control $\Lambda(\mathtt{Z})$. By independence of the $\mathtt{Z}_i$,

$$\mathrm{Var}(\Lambda(\mathtt{Z})) = \mathbb{E}\left[\|\Lambda(\mathtt{Z}) - \mathbb{E}[\Lambda(\mathtt{Z})]\|_{\mathrm{F}}^2\right] = \frac{1}{n}\mathbb{E}\left[\|\Lambda(\mathtt{Z}_1) - \mathbb{E}[\Lambda(\mathtt{Z}_1)]\|_{\mathrm{F}}^2\right]$$
$$\leqslant \frac{\mathrm{c}}{n}\mathbb{E}\left[\|\Lambda(\mathtt{Z}_1)\|_{\mathrm{F}}^2\right] \leqslant \frac{\mathrm{c}p}{n}\mathbb{E}\left[\|\mathrm{Var}(X_{\mathtt{Z}_1} \mid \mathtt{Z}_1)\|_{\mathrm{op}}^2\right].$$

Moreover,

$$\|\mathrm{Var}(X_{\mathtt{Z}_1} \mid \mathtt{Z}_1)\|_{\mathrm{op}} = \sup_{\|\mathbf{u}\|_2 = 1} \mathrm{Var}(\mathbf{u}^\top X_{\mathtt{Z}_1} \mid \mathtt{Z}_1) \leqslant \mathrm{c} \mathtt{L}_{\mathtt{x}}^2,$$

by (16). Therefore

$$\mathrm{Var}(\Lambda(\mathtt{Z})) \leqslant \mathrm{c} \frac{p}{n} \mathtt{L}_{\mathtt{x}}^4.$$

We now control $\Omega(\mathtt{Z})$. The same independence argument gives

$$\mathrm{Var}(\Omega(\mathtt{Z})) \leqslant \frac{\mathrm{c}}{n}\mathbb{E}\left[\|\mathrm{Cov}(X_{\mathtt{Z}_1}, Y_{\mathtt{Z}_1} \mid \mathtt{Z}_1)\|_2^2\right]$$
$$= \frac{\mathrm{c}}{n}\mathbb{E}\left[\left(\sup_{\|\mathbf{u}\|_2 = 1} \mathrm{Cov}(\mathbf{u}^\top X_{\mathtt{Z}_1}, Y_{\mathtt{Z}_1} \mid \mathtt{Z}_1)\right)^2\right].$$

By Cauchy–Schwarz, (16), and (17),

$$\mathrm{Cov}(\mathbf{u}^\top X_{\mathtt{Z}_1}, Y_{\mathtt{Z}_1} \mid \mathtt{Z}_1)^2 \leqslant \mathrm{c} \mathtt{L}_{\mathtt{x}}^2 \mathtt{L}_{\mathtt{y}}^2$$

uniformly over $\|\mathbf{u}\|_2 = 1$. Hence

$$\mathrm{Var}(\Omega(\mathtt{Z})) \leqslant \mathrm{c} \frac{1}{n} \mathtt{L}_{\mathtt{x}}^2 \mathtt{L}_{\mathtt{y}}^2.$$

□

Concrete examples. We now present several augmentation schemes for which **H**4 can be checked explicitly. These examples are meant to illustrate representative mechanisms used in practice, while Proposition 2 applies more broadly. For simplicity, all three examples are written in the identity feature map setting, namely $p(n) = d(n)$ and $\varphi(x) = x$ for all $x \in \mathbb{R}^{d(n)}$. This choice is made to keep the expressions of $\mu_{\mathtt{x}}$, $\mu_{\mathtt{y}}$, Λ and Ω explicit; the same arguments extend to non-identity feature maps whenever the assumptions of Proposition 2 are satisfied.

Example 1 (Heteroskedastic noise injection)**.** *Fix $k \in \mathbb{N}$, independent of n, and let η be a centered random vector in $\mathbb{R}^k$, independent of $\mathtt{Z}$, such that*

$$\mathbb{E}\left[\eta\eta^\top\right] = \mathrm{I}_k \ .$$

Let

$$s_{\mathtt{x}} : \mathbb{R}^{d(n)} \times \mathbb{R} \to \mathbb{R}^{d(n)\times k} \ , \qquad\qquad s_{\mathtt{y}} : \mathbb{R}^{d(n)} \times \mathbb{R} \to \mathbb{R}^{1\times k}$$

be maps satisfying, for some constants $\mathtt{L}_s, B > 0$ independent of n,

$$\begin{aligned} \|s_{\mathtt{x}}(z) - s_{\mathtt{x}}(z')\|_{\mathrm{F}} + \|s_{\mathtt{y}}(z) - s_{\mathtt{y}}(z')\|_2 &\leqslant \mathtt{L}_s\|z - z'\|_2 \ , \\ \|s_{\mathtt{x}}(z)\|_{\mathrm{op}} + \|s_{\mathtt{y}}(z)\|_2 &\leqslant B \end{aligned}$$

for all $z, z' \in \mathbb{R}^{d(n)} \times \mathbb{R}$. For $z = (x, y) \in \mathbb{R}^{d(n)} \times \mathbb{R}$ and deterministic $\boldsymbol{\eta} \in \mathbb{R}^k$, define

$$\tau_{\mathtt{x}}(z, \boldsymbol{\eta}) = x + s_{\mathtt{x}}(z)\boldsymbol{\eta} \ , \qquad\qquad \tau_{\mathtt{y}}(z, \boldsymbol{\eta}) = y + s_{\mathtt{y}}(z)\boldsymbol{\eta} \ ,$$

so that the random augmentation is obtained by evaluating these maps at $\boldsymbol{\eta} = \eta$. Since η is centered with identity covariance, we get

$$\begin{aligned} \mu_{\mathtt{x}}(z) &= x \ , & \mu_{\mathtt{y}}(z) &= y \ , \\ \Lambda(z) &= s_{\mathtt{x}}(z)s_{\mathtt{x}}(z)^\top \ , & \Omega(z) &= s_{\mathtt{x}}(z)s_{\mathtt{y}}(z)^\top \ . \end{aligned}$$

The required regularity bounds follow directly from those of $s_{\mathtt{x}}$ and $s_{\mathtt{y}}$. Indeed,

$$\begin{aligned} \|\Lambda(z) - \Lambda(z')\|_{\mathrm{F}} &\leqslant \|s_{\mathtt{x}}(z)\|_{\mathrm{op}}\|s_{\mathtt{x}}(z) - s_{\mathtt{x}}(z')\|_{\mathrm{F}} + \|s_{\mathtt{x}}(z')\|_{\mathrm{op}}\|s_{\mathtt{x}}(z) - s_{\mathtt{x}}(z')\|_{\mathrm{F}} \\ &\leqslant 2B\mathtt{L}_s\|z - z'\|_2 \ , \end{aligned}$$

and similarly $\|\Omega(z) - \Omega(z')\|_2 \leqslant 2B\mathtt{L}_s\|z - z'\|_2$. Moreover, since k is fixed,

$$\|\Lambda(z)\|_{\mathrm{F}} \leqslant \sqrt{k}\, B^2 \ , \qquad \|\Omega(z)\|_2 \leqslant B^2 \ ,$$

*and therefore $\mathrm{Var}(\Lambda(\mathtt{Z}_1)) + \mathrm{Var}(\Omega(\mathtt{Z}_1)) \leqslant \mathrm{c}$. Therefore **H**4 holds.*

Example 2 (Salt-and-pepper noise injection)**.** *We consider the coordinatewise salt-and-pepper scheme introduced in the main body. Let $\mathfrak{m} : \mathbb{R}^{d(n)} \times \mathbb{R} \to [0, 1]$ and $s : \mathbb{R}^{d(n)} \times \mathbb{R} \to \mathbb{R}^{d(n)\times d(n)}$. For $z = (x, y) \in \mathbb{R}^{d(n)} \times \mathbb{R}$ and deterministic $\boldsymbol{\eta} = (\{u_i\}_{i=1}^{d(n)}, \boldsymbol{\eta}_2)$, with $\boldsymbol{\eta}_2 \in \mathbb{R}^{d(n)}$, let $m_{z,\boldsymbol{\eta}}$ be the vector with i-th component $\mathbb{1}\{u_i \leqslant \mathfrak{m}(z)\}$. We define*

$$\begin{aligned} \tau_{\mathtt{x}}(z, \boldsymbol{\eta}) &= x \odot m_{z,\boldsymbol{\eta}} + s(z)\left(\boldsymbol{\eta}_2 - \boldsymbol{\eta}_2 \odot m_{z,\boldsymbol{\eta}}\right) \ , \\ \tau_{\mathtt{y}}(z, \boldsymbol{\eta}) &= y \ , \end{aligned}$$

and obtain the random augmentation by evaluating these maps at $\boldsymbol{\eta} = \eta$. Here $\eta = (\{U_i\}_{i=1}^{d(n)}, \eta_2)$ is independent of $\mathtt{Z}$, the variables $\{U_i\}_{i=1}^{d(n)}$ are i.i.d. with distribution $\mathrm{Unif}([0, 1])$, and $\eta_2 \in \mathbb{R}^{d(n)}$ is centered, independent of $\{U_i\}_{i=1}^{d(n)}$, with

$$\mathbb{E}\left[\eta_2\eta_2^\top\right] = \mathrm{I}_{d(n)} \ .$$

Writing $m_z = m_{z,\eta}$, we have $\mathbb{E}[m_z] = \mathfrak{m}(z)\mathbf{1}_{d(n)}$ and $\eta_2 - \eta_2 \odot m_z$ is centered conditionally on m_z. Hence

$$\begin{aligned} \mu_{\mathtt{x}}(z) &= \mathfrak{m}(z)x \ , & \mu_{\mathtt{y}}(z) &= y \ , \\ \Lambda(z) &= \{1 - \mathfrak{m}(z)\}\left[\mathfrak{m}(z)\mathrm{Diag}(xx^\top) + s(z)s(z)^\top\right] \ , & \Omega(z) &= 0 \ . \end{aligned}$$

Indeed, by independence of m_z and η_2,

$$\mathbb{E}\left[\left(\eta_2 - \eta_2 \odot m_z\right)\left(\eta_2 - \eta_2 \odot m_z\right)^\top\right] = \{1 - \mathfrak{m}(z)\}\mathrm{I}_{d(n)} \ ,$$

which gives the expression of $\Lambda(z)$, while $\Omega(z) = 0$ because $\tau_{\mathtt{y}}(z, \eta) = y$ is deterministic conditionally on z. Assume moreover that $\mathfrak{m}$ and s are Lipschitz continuous, with Lipschitz constants uniformly bounded in n, and that

$$\sup_{n\in\mathbb{N}} \sup_{z\in\mathbb{R}^{d(n)}\times\mathbb{R}} \|s(z)\|_{\mathrm{op}} < \infty \, .$$

If the law of $\mathtt{Z}_1$ is supported in a Euclidean ball of radius independent of n, then the maps $\mu_{\mathtt{x}}$, $\mu_{\mathtt{y}}$, Λ and Ω are bounded and Lipschitz on this support, with constants uniformly bounded in n. Therefore

$$\mathrm{Var}(\Lambda(\mathtt{Z}_1)) + \mathrm{Var}(\Omega(\mathtt{Z}_1)) \leqslant \mathrm{c} \, ,$$

*and **H**4 holds for this scheme.*

Example 3 (Label-aware mixtures of augmentations)**.** *Fix $m \in \mathbb{N}$, independent of n. For each $j \in \{1, \dots, m\}$, let η_j be an augmentation variable taking values in a space E_j, and let*

$$\tau_{\mathtt{x},j} : \mathbb{R}^{d(n)} \times \mathbb{R} \times \mathsf{E}_j \to \mathbb{R}^{d(n)} \, , \qquad \tau_{\mathtt{y},j} : \mathbb{R}^{d(n)} \times \mathbb{R} \times \mathsf{E}_j \to \mathbb{R}$$

be an arbitrary augmentation mechanism. The variables $\eta_1, \dots, \eta_m$ are assumed independent of $\mathtt{Z}$. Let $\pi_1, \dots, \pi_m : \mathbb{R}^{d(n)} \times \mathbb{R} \to [0, 1]$ be Lipschitz continuous maps satisfying

$$\sum_{j=1}^{m} \pi_j(z) = 1$$

for all $z \in \mathbb{R}^{d(n)} \times \mathbb{R}$. Let V be uniformly distributed on $[0, 1]$, independent of $\mathtt{Z}$ and of $\eta_1, \dots, \eta_m$. For deterministic $v \in [0, 1]$, define the selected augmentation index

$$J(z, v) = \min \left\{ j \in \{1, \dots, m\} : \sum_{s=1}^{j} \pi_s(z) \geqslant v \right\} \, .$$

For $z = (x, y) \in \mathbb{R}^{d(n)} \times \mathbb{R}$ and deterministic $\boldsymbol{\eta} = (v, \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_m)$, define

$$\tau_{\mathtt{x}}(z, \boldsymbol{\eta}) = \tau_{\mathtt{x},J(z,v)}(z, \boldsymbol{\eta}_{J(z,v)}) \, ,$$
$$\tau_{\mathtt{y}}(z, \boldsymbol{\eta}) = \tau_{\mathtt{y},J(z,v)}(z, \boldsymbol{\eta}_{J(z,v)}) \, ,$$

and obtain the random augmentation by evaluating these maps at $\boldsymbol{\eta} = (V, \eta_1, \dots, \eta_m)$. For each component, define

$$\mu_{\mathtt{x},j}(z) = \mathbb{E}\left[\tau_{\mathtt{x},j}(z, \eta_j)\right] \, , \qquad \mu_{\mathtt{y},j}(z) = \mathbb{E}\left[\tau_{\mathtt{y},j}(z, \eta_j)\right] \, ,$$
$$\Lambda_j(z) = \mathrm{Cov}\left(\tau_{\mathtt{x},j}(z, \eta_j)\right) \, , \quad \Omega_j(z) = \mathrm{Cov}\left(\tau_{\mathtt{x},j}(z, \eta_j), \tau_{\mathtt{y},j}(z, \eta_j)\right) \, .$$

Since $\mathbb{P}(J(z, V) = j) = \pi_j(z)$, the resulting quantities are

$$\mu_{\mathtt{x}}(z) = \sum_{j=1}^{m} \pi_j(z)\mu_{\mathtt{x},j}(z) \, , \qquad \mu_{\mathtt{y}}(z) = \sum_{j=1}^{m} \pi_j(z)\mu_{\mathtt{y},j}(z) \, ,$$

$$\Lambda(z) = \sum_{j=1}^{m} \pi_j(z) \left\{\Lambda_j(z) + \mu_{\mathtt{x},j}(z)\mu_{\mathtt{x},j}(z)^\top\right\} - \mu_{\mathtt{x}}(z)\mu_{\mathtt{x}}(z)^\top \, ,$$

$$\Omega(z) = \sum_{j=1}^{m} \pi_j(z) \left\{\Omega_j(z) + \mu_{\mathtt{x},j}(z)\mu_{\mathtt{y},j}(z)\right\} - \mu_{\mathtt{x}}(z)\mu_{\mathtt{y}}(z) \, .$$

Assume that the component-level maps $\mu_{\mathtt{x},j}, \mu_{\mathtt{y},j}, \Lambda_j, \Omega_j$ are bounded and Lipschitz with constants uniformly bounded in j and in n. Since m is fixed and the weights π_j are Lipschitz, the maps $\mu_{\mathtt{x}}, \mu_{\mathtt{y}}, \Lambda$ and Ω are also bounded and Lipschitz with constants uniformly bounded in n. Therefore

$$\mathrm{Var}(\Lambda(\mathtt{Z}_1)) + \mathrm{Var}(\Omega(\mathtt{Z}_1)) \leqslant \mathrm{c} \, ,$$

*and **H**4 holds.*

The previous proposition and examples should be viewed as elementary building blocks rather than isolated constructions. They show that the regularity requirements in **H**4 are stable under several common augmentation mechanisms, including sample-dependent noise injection, masking, and finite mixtures of label-aware transformations. Consequently, the framework covers a broad family of practical data-augmentation pipelines, including schemes whose strength or structure adapts to the local geometry of the data or to label-dependent invariances.

C Technical preliminaries on concentration of measure

We work throughout the appendix under the *Lipschitz concentration* framework of [LC18, LC20], which will serve as the main probabilistic tool in the proofs of Theorem 1 and Proposition 1 and is defined as follows:

Definition 2. *A random variable $Z \in \mathbb{R}^d$ is said to be* Lipschitz concentrated *with parameter κ, if for all 1-Lispchitz function $f : \mathbb{R}^d \to \mathbb{R}$, $f(Z)$ is a κ-sub-Gaussian random variable.*

In particular, the concentration property assumed for $\mathtt{Z}$ in **H**1 simply restate as saying that the columns of $\mathtt{Z}$ are 1-Lispchitz concentrated. Furthermore, [LC18] section 2.2.2. shows that such assumption implies that the full matrix $\mathtt{Z}$ is 1-Lipschitz concentrated, which will be extensively used in the next sections. We then collect several technical results that will be repeatedly used in the sequel. Section C.1 introduces arguments that disentangle the dependence between training samples, most notably we generalize the standard 'leave-one-out' argument of [Cho22] to a 'leave-two-out' argument, and establish a set of concentration bounds for large random matrices. Section C.2 focuses on the concentration of random quadratic forms; in particular, we restate the Hanson–Wright inequality [Ada15] in a setting where both the covariate and the quadratic form may be random. Finally, Section C.3 shows that random variables satisfying Definition 2 retain satisfying concentration properties after composition with certain non-Lipschitz mappings. All our concentration results will be stated in the form of a stochastic domination result, introduced in Definition 1.

C.1 Concentration of large random matrices: Useful lemmas

We begin by recalling a standard fact regarding the spectral properties of the covariance matrix of a Lipschitz concentrated random vector. This simple observation will be useful for bounding operator norms of covariance matrices throughout the paper, thus allowing us to derive bounds that depend only on the Lipschitz concentration parameter of the data.

Lemma 1 (Operator norm bound for the covariance matrix of a Lipschitz concentrated vector)**.** *Let Z be a centered random vector in $\mathbb{R}^d$ which is κ-Lipschitz concentrated. Then,*

$$\|\mathbb{E}[ZZ^\top]\|_{\mathrm{op}} \leqslant \kappa^2 \, .$$

Proof. By definition of the operator norm, we write,

$$\|\mathbb{E}[ZZ^\top]\|_{\mathrm{op}} = \sup_{\|u\|_2=1} u^\top \mathbb{E}[ZZ^\top]u = \sup_{\|u\|_2=1} \mathrm{Var}(u^\top Z) \, .$$

For any $u \in \mathbb{R}^d$, the mapping $z \mapsto u^\top z$ is $\|u\|_2$-Lipschitz, thus $u^\top Z$ is $\kappa\|u\|_2$-sub-Gaussian from the Lipschitz concentration assumption. In particular, for unit vectors u, we have $\mathrm{Var}(u^\top Z) \leqslant \kappa^2$. The claimed bound follows. □

A more general result holds for any Lipschitz concentrated random matrix,

Lemma 2 (Spectral norm with Lipschitz–concentrated rows)**.** *Let $A \in \mathbb{R}^{p\times d}$ have independent rows $A_1, \ldots, A_p \in \mathbb{R}^d$. Assume each row is κ–Lipschitz concentrated. Then there exists $\mathrm{c} > 0$ such that for all $s > 0$,*

$$\mathbb{P}\Big(\|A - \mathbb{E}[A]\|_{\mathrm{op}} \;\geqslant\; \mathrm{a}\,\kappa\,(\sqrt{p} + \sqrt{d}) \;+\; s \Big) \;\leqslant 2\,\exp(-\,\mathrm{c}\,s^2/\kappa^2) \, , \tag{18}$$

and,

$$\mathbb{P}\Big(\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}} \geqslant 2\mathrm{a}\,\kappa\,(\sqrt{p} + \sqrt{d})\Big) \;\leqslant 2\,\exp(-\,\mathrm{c}\mathrm{a}_1^2(p+d)) \, . \tag{19}$$

with a *defined as,*

$$\mathrm{a} = \sqrt{2\log(9)} \, . \tag{20}$$

Furthermore, for all $k \in \mathbb{N}^*$, and for three constants $\mathrm{c}_k, \mathrm{c}'_k$, we have,

$$\mathbb{E}\left[\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}}^k\right] \leqslant \mathrm{c}_k\kappa^k(\sqrt{p} + \sqrt{d} + 1)^k + \mathrm{c}'_k\kappa^{k-1}(\sqrt{p} + \sqrt{d})^{k-1} \, . \tag{21}$$

Proof. The proof follows the standard ε–net argument (see, e.g., Lemma 5.2 and Lemma 5.4 of [Ver18]). We consider ε-nets of the unit spheres $\mathbb{S}^{p-1}$ and $\mathbb{S}^{d-1}$, i.e., sets $\mathsf{N} \subset \mathbb{S}^{p-1}$ and $\mathsf{M} \subset \mathbb{S}^{d-1}$ such that for any $x \in \mathbb{S}^{p-1}$, there exists $x' \in \mathsf{N}$ with $\|x - x'\|_2 \leqslant \varepsilon$, and similarly for M. For $\varepsilon = 1/4$, Corollary 4.2.13 of [Ver18] guarantees the existence of ε–nets with cardinalities:

$$|\mathsf{N}| \leqslant (1 + 2/\varepsilon)^p \leqslant 9^p, \qquad |\mathsf{M}| \leqslant (1 + 2/\varepsilon)^d \leqslant 9^d \,. \tag{22}$$

For any matrix $B \in \mathbb{R}^{p \times d}$, the operator norm can be approximated by the maximum over the ε-nets:

$$\|B\|_{\mathrm{op}} = \max_{x \in \mathbb{S}^{p-1},\, y \in \mathbb{S}^{d-1}} x^\top B y \leqslant \frac{1}{1 - 2\varepsilon} \max_{x \in \mathsf{N},\, y \in \mathsf{M}} x^\top B y \leqslant 2 \max_{x \in \mathsf{N},\, y \in \mathsf{M}} x^\top B y \,. \tag{23}$$

For any $x \in \mathsf{N}$ and $y \in \mathsf{M}$, define

$$S_{x,y} := x^\top (A - \mathbb{E}[A])\, y = \sum_{i=1}^{p} x_i \, \langle A_i - \mathbb{E}[A_i], y \rangle \,.$$

Since the map $r \mapsto \langle r, y \rangle$ is 1–Lipschitz on $\mathbb{R}^d$, and each row A_i is κ–Lipschitz concentrated, each summand $X_i := \langle A_i - \mathbb{E}[A_i], y \rangle$ is zero-mean sub-Gaussian with parameter κ. Since the $\{A_i\}_{i=1}^p$ are independent, $\{X_i\}_{i=1}^p$ are also independent, and therefore $S_{x,y} = \sum_i x_i X_i$ is sub-Gaussian with parameter $\kappa \|x\|_2 = \kappa$. Therefore, for any $t \geqslant 0$:

$$\mathbb{P}\big(|S_{x,y}| \geqslant t\big) \leqslant 2 \exp(-t^2/(2\kappa^2)) \,.$$

By the union bound over $\mathsf{N} \times \mathsf{M}$ and using (22):

$$\mathbb{P}\left(\max_{x \in \mathsf{N},\, y \in \mathsf{M}} |S_{x,y}| \geqslant t \right) \leqslant 2 \cdot 9^{p+d} \exp\left(-\frac{t^2}{2\kappa^2} \right) \,.$$

Setting $t = \sqrt{2\log(9)}\,\kappa\,(\sqrt{p} + \sqrt{d}) + s$ and using $\log(9^{p+d}) = (p+d)\log(9) \leqslant (\sqrt{p} + \sqrt{d})^2 \log(9)$:

$$\mathbb{P}\left(\max_{x \in \mathsf{N},\, y \in \mathsf{M}} |S_{x,y}| \geqslant \sqrt{2\log(9)}\,\kappa\,(\sqrt{p} + \sqrt{d}) + s \right) \leqslant 2 \exp\left(-\frac{s^2}{2\kappa^2} \right) \,.$$

Using (23) with $B = A - \mathbb{E}[A]$ and defining $\mathrm{a}_1 = \sqrt{2\log(9)}$ completes the proof of (18).
The second equation (19) follows by setting $s = \mathrm{a}_1\,\kappa\,(\sqrt{p} + \sqrt{d})$ and using $(\sqrt{p} + \sqrt{d})^2 \geqslant (p + d)$.
For the moment bound (21), we use the integral representation:

$$\mathbb{E}\left[\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}}^k \right] = \int_0^\infty k t^{k-1} \mathbb{P}\left(\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}} \geqslant t \right) dt \,.$$

We split the integral at $t_0 = 2\mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d})$:

$$\begin{aligned} \mathbb{E}\left[\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}}^k \right] &= \int_0^{t_0} k t^{k-1} dt + \int_{t_0}^\infty k t^{k-1} \mathbb{P}\left(\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}} \geqslant t \right) dt \\ &= t_0^k + \int_{t_0}^\infty k t^{k-1} \mathbb{P}\left(\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}} \geqslant t \right) dt \,. \end{aligned}$$

For the second integral, we use the concentration bound from (18) with $s = t - \mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d})$:

$$\begin{aligned} \int_{t_0}^\infty k t^{k-1} \mathbb{P}\left(\|A - \mathbb{E}\,[A]\,\|_{\mathrm{op}} \geqslant t \right) dt &\leqslant 2k \int_{t_0}^\infty t^{k-1} \exp\left(-\frac{(t - \mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d}))^2}{2\kappa^2} \right) dt \\ &= 2k \int_0^\infty (s + \mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d}))^{k-1} \exp\left(-\frac{s^2}{2\kappa^2} \right) ds \,, \end{aligned}$$

where we made the substitution $s = t - \mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d})$. Using $(a+b)^{k-1} \leqslant 2^{k-2}(a^{k-1} + b^{k-1})$ for $a, b \geqslant 0$:

$$\begin{aligned} &\leqslant 2^{k-1} k \left[(\mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d}))^{k-1} \int_0^\infty \exp\left(-\frac{s^2}{2\kappa^2} \right) ds + \int_0^\infty s^{k-1} \exp\left(-\frac{s^2}{2\kappa^2} \right) ds \right] \\ &= 2^{k-1} k \left[(\mathrm{a}_1\kappa(\sqrt{p} + \sqrt{d}))^{k-1} \Gamma(1)\kappa + \left(\sqrt{2}\kappa \right)^{k-1} \Gamma\left(\frac{k}{2} \right) \right] \,. \end{aligned}$$

Combining both terms and using that $t_0^k = (2\mathrm{a}_1\kappa(\sqrt{p}+\sqrt{d}))^k$:

$$\begin{aligned}\mathbb{E}\left[\|A-\mathbb{E}[A]\|_{\mathrm{op}}^k\right] &\leqslant (2\mathrm{a}_1\kappa(\sqrt{p}+\sqrt{d}))^k + 2^{k-1}k\sqrt{\pi/2}\kappa(\mathrm{a}_1\kappa(\sqrt{p}+\sqrt{d}))^{k-1} \\ &\quad + 2^{k-1}k\left(\sqrt{2}\kappa\right)^{k-1}\Gamma\left(\frac{k}{2}\right) \\ &\leqslant \mathrm{c}_k\kappa^k(\sqrt{p}+\sqrt{d}+1)^k + \mathrm{c}'_k\kappa^{k-1}(\sqrt{p}+\sqrt{d})^{k-1}\ ,\end{aligned}$$

for appropriate constants $\mathrm{c}_k, \mathrm{c}'_k > 0$ that depend only on k. □

Now, recall the definition of the variance of a random matrix. For any $A \in \mathbb{R}^{p\times p}$,

$$\mathrm{Var}(A) = \mathbb{E}\left[\|A-\mathbb{E}[A]\|_{\mathrm{F}}^2\right]\ . \tag{24}$$

Note that this definition is consistent with the one of the variance of a random variable living in a general Hilbert space. We prove the following standard result about the variance of a product of random matrices, although it holds in any Hilbert space, we state it here only in the case of $A \in \mathbb{R}^{p\times p}$.

Lemma 3 (Variance of the product of random matrices)**.** *Let $A_1, A_2 \in \mathbb{R}^{p\times p}$ be two random matrices. Then,*

$$\begin{aligned}&\mathrm{Var}(A_1A_2) \leqslant 2\|\|A_1\|_{\mathrm{op}}\|_\infty^2\mathrm{Var}(A_2) + 2\|\mathbb{E}[A_2]\|_{\mathrm{op}}^2\mathrm{Var}(A_1), \\ &\mathrm{Var}(A_1A_2) \leqslant 2\|\|A_2\|_{\mathrm{op}}\|_\infty^2\mathrm{Var}(A_1) + 2\|\mathbb{E}[A_1]\|_{\mathrm{op}}^2\mathrm{Var}(A_2), \\ \textit{and}\quad &\mathrm{Var}(A_1A_2A_1) \leqslant 2\|\|A_1\|_{\mathrm{op}}\|_\infty^4\mathrm{Var}(A_2) + 8\|\mathbb{E}[A_1]\|_{\mathrm{op}}^2\|\|A_1\|_{\mathrm{op}}\|_\infty^2\mathrm{Var}(A_2)\ .\end{aligned}$$

Proof. We prove the first inequality, the second one follows from the same arguments. We have that,

$$\mathrm{Var}(A_1A_2) = \mathrm{Var}(A_1\{A_2-\mathbb{E}[A_2]\} + A_1\mathbb{E}[A_2]) \leqslant 2\mathrm{Var}(A_1\{A_2-\mathbb{E}[A_2]\}) + 2\mathrm{Var}(A_1\mathbb{E}[A_2])\ .$$

Using $\mathrm{Var}(A) \leqslant \mathbb{E}\left[\|A\|_{\mathrm{F}}^2\right]$, it follows that,

$$\mathrm{Var}(A_1\{A_2-\mathbb{E}[A_2]\}) \leqslant \mathbb{E}\left[\|A_1\{A_2-\mathbb{E}[A_2]\}\|_{\mathrm{F}}^2\right] \leqslant \|\|A_1\|_{\mathrm{op}}\|_\infty^2\mathrm{Var}(A_2)\ .$$

Respectively, and by definition of Var, we have,

$$\mathrm{Var}(A_1\mathbb{E}[A_2]) = \mathbb{E}\left[\|\{A_1-\mathbb{E}[A_1]\}\mathbb{E}[A_2]\|_{\mathrm{F}}^2\right] \leqslant \|\mathbb{E}[A_2]\|_{\mathrm{op}}^2\mathrm{Var}(A_1)\ .$$

The two first inequalities follow. We proove the last inequality in a similar way, indeed we have,

$$\mathrm{Var}(A_1A_2A_1) \leqslant 2\mathrm{Var}(A_1\{A_2-\mathbb{E}[A_2]\}A_1) + 2\mathrm{Var}(A_1\mathbb{E}[A_2]\,A_1)\ ,$$

Using the previously derived upper bounds, we have,

$$\begin{aligned}\mathrm{Var}(A_1\mathbb{E}[A_2]\,A_1) &\leqslant 2\|\|A_1\|_{\mathrm{op}}\|_\infty^2\mathrm{Var}\left(\mathbb{E}[A_2]\,A_1\right) + 2\|\mathbb{E}[A_1]\|_{\mathrm{op}}^2\|\mathbb{E}[A_2]\|_{\mathrm{op}}^2\mathrm{Var}(A_1) \\ &\leqslant 4\|\|A_1\|_{\mathrm{op}}\|_\infty^2\|\mathbb{E}[A_2]\|_{\mathrm{op}}^2\mathrm{Var}(A_1)\end{aligned}$$

in addition,

$$\mathrm{Var}(A_1\{A_2-\mathbb{E}[A_2]\}A_1) \leqslant \mathbb{E}\left[\|A_1\{A_2-\mathbb{E}[A_2]\}A_1\|_{\mathrm{F}}^2\right] \leqslant \|\|A_1\|_{\mathrm{op}}\|_\infty^4\mathrm{Var}(A_2)\ .$$

The statement follows. □

C.2 Concentration of random quadratic forms

Moving on, we now restate a few usual results about the concentration of quadratic forms in the context of stochastic domination Definition 1 and Lipschitz concentration Definition 2. We start by highlighting a few standard properties of stochastic domination, and how it relates to non-asymptotic concentration bounds. Our first result shows the additivity and multiplicative properties of Definition 1.

Lemma 4 (Stability of the stochastic domination)**.** *Let $(U_n)_{n\in\mathbb{N}}, (V_n)_{n\in\mathbb{N}}, (U'_n)_{n\in\mathbb{N}}$, and $(V'_n)_{n\in\mathbb{N}}$ be four sequences of non-negative random variables. If $U_n \prec_s V_n$ and $U'_n \prec_s V'_n$, then it holds,*

$$U_n + U'_n \prec_s V_n + V'_n\ ,\quad \textit{and}\quad U_nU'_n \prec_s V_nV'_n\ .$$

Proof. Fix $\delta > 0$ and $D > 0$. From $U_n \prec_s V_n$ and $U'_n \prec_s V'_n$, by definition of stochastic domination, for any choice of $\delta > 0$ and $D > 0$ there exists $N(\delta, D)$ such that for all $n \geqslant N(\delta, D)$,

$$\mathbb{P}\big(U_n > n^\delta V_n\big) \leqslant n^{-D} \quad \text{and} \quad \mathbb{P}\big(U'_n > n^\delta V'_n\big) \leqslant n^{-D}.$$

Additivity. Applying the union-bound, we have, for $n \geqslant N(\delta, D+1)$,

$$\mathbb{P}\big(U_n + U'_n > n^\delta (V_n + V'_n)\big) \;\leqslant\; \mathbb{P}\big(U_n > n^\delta V_n\big) + \mathbb{P}\big(U'_n > n^\delta V'_n\big) \;\leqslant\; 2n^{-D+1} \leqslant n^{-D},$$

where the last inequality holds for sufficiently large n. Thus $U_n + U'_n \prec_s V_n + V'_n$.

Multiplicativity. Similarly as in the additive case, we have thanks to the union bound,

$$\mathbb{P}\big(U_n U'_n > n^\delta V_n V'_n\big) \leqslant \mathbb{P}\Big(U_n > n^{\delta/2} V_n\Big) + \mathbb{P}\Big(U'_n > n^{\delta/2} V'_n\Big) \leqslant 2n^{-(D+1)} \leqslant n^{-D},$$

for all sufficiently large n. Hence $U_n U'_n \prec_s V_n V'_n$.

Both claims hold for arbitrary $\delta > 0$ and $D > 0$, which completes the proof. □

We now clarify the connection between stochastic domination and classical concentration of measure results. To this end, introduce the following definition, which can be seen as an extension of the definition of a sub-Gaussian random variable to sequences of random variables, such as in Definition 1.

Definition 3. *We say that a sequence of real random variables $(U_n)_{n\in\mathbb{N}}$ is a sub-Gaussian sequence with parameters $(\sigma_n)_{n\in\mathbb{N}}$ if for there exists $\mathrm{c} \geqslant 0$ and $\lambda_0 > 0$ such that for any $n \in \mathbb{N}$ and $\lambda \leqslant \lambda_0$, $\mathbb{E}[\mathrm{e}^{\lambda U_n}] \leqslant \exp(\mathrm{c}\lambda^2\sigma_n^2)$. If $(U_n)_{n\in\mathbb{N}}$ is complex valued, we say similarly that it is a sub-Gaussian sequence with parameters $(\sigma_n)_{n\in\mathbb{N}}$ if its real and imaginary parts are sub-Gaussian sequences with parameters $(\sigma_n)_{n\in\mathbb{N}}$.*

We observe that several common tail bounds imply stochastic domination in the appropriate regimes.

Lemma 5 (Stochastic domination from sub-Gaussian and sub-Exponential concentration)**.** *Let $(U_n)_n$ be a sequence of complex random variables. Suppose there exist constants $\mathrm{c}_1, \mathrm{c}_2, \alpha > 0$ and a sequence $(\sigma_n)_n$ of positive real numbers such that, for all $t > 0$,*

$$\mathbb{P}\left(|U_n| > t\right) \leqslant \mathrm{c}_1 \exp\left(-\mathrm{c}_2 \frac{t^\alpha}{\sigma_n^\alpha}\right) .$$

Then, it follows that

$$|U_n| \prec_s \sigma_n \, .$$

Proof. Let $\varepsilon, D > 0$. For all $n \geqslant 1$, we may write

$$\mathbb{P}\left(|X_n| > n^\varepsilon \sigma_n\right) \leqslant \mathrm{c}_1 \exp\left(-\mathrm{c}_2 \frac{(n^\varepsilon \sigma_n)^\alpha}{\sigma_n^\alpha}\right) = \mathrm{c}_1 \exp\left(-\mathrm{c}_2 n^{\alpha\varepsilon}\right).$$

The exponential decay ensures that, for sufficiently large n, $\mathrm{c}_1 \exp\left(-\mathrm{c}_2 n^{\alpha\varepsilon}\right) \leqslant n^{-D}$. Therefore, the conclusion follows by the definition of stochastic domination. □

To conclude this subsection, we provide a version of the Hanson-Wright inequality in the framework of stochastic domination. To this end, we first recall the usual statement of the Hanson-Wright inequality:

Lemma 6 (Hanson-Wright inequality)**.** *Let $(Z_n)_n$ be a sequence of centered d dimensional random vector such that for all $n \in \mathbb{N}$, Z_n is σ_n-Lipschitz concentrated. For a sequence of deterministic matrix $(\mathbf{A}_n)_n$, it holds for all $t > 0$,*

$$\mathbb{P}\left(\left|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]\right| > t\right) \leqslant 2\exp\left(-c\min\left\{\frac{t^2}{\sigma_n^4 \|\mathbf{A}_n\|_F^2}, \frac{t}{\sigma_n^2 \|\mathbf{A}_n\|_{\mathrm{op}}}\right\}\right) , \tag{25}$$

Stated as a stochastic domination result, we have,

$$\left|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]\right| \prec_s \sigma_n^2 \|\mathbf{A}_n\|_F \, .$$

And, the moments of such quadratic forms are bounded by,

$$\mathbb{E}\left[\left|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]\right|^k\right] \leqslant \mathrm{c}_k \sigma_n^{2k} \|\mathbf{A}_n\|_F^k \, .$$

Where c, $(c_k)_{k\in\mathbb{N}} > 0$ are universal constants, that don't depend on n, Z_n nor $\mathbf{A}_n$.

Proof. The first part of the theorem is a direct application of Theorem 2.4. of [Ada15]. The second part follows from Lemma 5. We prove the third part by writting the moments in integral form as, and using (25),

$$\mathbb{E}\left[|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]|^k\right] = \int_0^\infty kt^{k-1}\mathbb{P}\left(\left|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]\right| > t\right) dt$$
$$\leqslant 2k \int_0^\infty t^{k-1} \exp\left(-c\min\left\{\frac{t^2}{\sigma_n^4\|\mathbf{A}_n\|_F^2}, \frac{t}{\sigma_n^2\|\mathbf{A}_n\|_{\mathrm{op}}}\right\}\right) dt \ .$$

Let $u = t/(\sigma_n^2\|\mathbf{A}_n\|_F)$ and $v = \|\mathbf{A}_n\|_{\mathrm{op}}/\|\mathbf{A}_n\|_F$. The integral becomes:

$$\leqslant 2k(\sigma_n^2\|\mathbf{A}_n\|_F)^k \int_0^\infty u^{k-1} \exp\left(-c\min\left\{u^2, \frac{u}{v}\right\}\right) du \ .$$

Splitting the integral into two parts, we have:

$$\int_0^\infty u^{k-1}e^{-c\min\{u^2,u/v\}}du \leqslant \int_0^\infty u^{k-1}e^{-cu^2}du + \int_0^\infty u^{k-1}e^{-cu/v}du = \frac{\Gamma(k/2)}{2c^{k/2}} + \frac{\Gamma(k)}{c^k} \ .$$

Where we have used $v < 1$. We can conclude that:

$$\mathbb{E}\left[|Z_n^\top \mathbf{A}_n Z_n - \mathbb{E}[Z_n^\top \mathbf{A}_n Z_n]|^k\right] \leqslant \mathrm{c}_k\sigma_n^{2k}\|\mathbf{A}_n\|_F^k \ ,$$

where $\mathrm{c}_k = 2k\left(\Gamma(k/2)/(2c^{k/2}) + \Gamma(k)/(c^k)\right)$. □

In what follows, it will also be necessary to address the scenario in which $\mathbf{A}_n$ is a random matrix that is independent of X_n. Specifically, we establish the following result:

Corollary 1 (Concentration of random quadratic forms)**.** *Consider a sequence of centered d-dimensional random vectors $(Z_n)_{n\in\mathbb{N}}$ that are σ_n-Lipschitz concentrated, and a sequence of random matrices $(A_n)_{n\in\mathbb{N}}$ that are κ_n-Lipschitz concentrated and independent of $(Z_n)_{n\in\mathbb{N}}$. Then, the deviation from its expectation is stochastically dominated by,*

$$|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n\right]| \prec_s \sigma_n^2\|A_n\|_{\mathrm{F}} + \kappa_n\|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}} \ .$$

And for $k \in \mathbb{N}$, its k-th moment satisfies

$$\mathbb{E}\left[\left|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n\right]\right|^k\right] \leqslant \mathrm{c}_k\left(\sigma_n^{2k}\mathbb{E}\left[\|A_n\|_{\mathrm{F}}^k\right] + \kappa_n^k\|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}}^k\right) \ ,$$

where $(\mathrm{c}_k)_{k\in\mathbb{N}} > 0$ are universal constants.

Proof. We establish the result by decomposing the deviation into two components and analyzing each separately. First, we write the triangular inequality:

$$|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n\right]| \leqslant |Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]| + |\mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right] - \mathbb{E}\left[Z_n^\top A_n Z_n\right]| \ . \tag{26}$$

By applying Lemma 6 conditionally on A_n, we obtain the high probability bound:

$$\mathbb{P}\left(|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]| \geqslant t \mid A_n\right) \leqslant 2\exp\left(-\mathrm{c}\min\left\{\frac{t^2}{2\sigma_n^4\|A_n\|_{\mathrm{F}}^2}, \frac{t}{\sigma_n\|A_n\|_{\mathrm{op}}}\right\}\right) \ .$$

Since $\|A_n\|_{\mathrm{op}} \leqslant \|A_n\|_{\mathrm{F}}$ almost surely, it follows from the definition of stochastic domination that:

$$|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]| \prec_s \sigma_n^2\|A_n\|_{\mathrm{F}} \ .$$

Now turning to the second term in (26), we express the conditional expectation $\mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]$ as a function of A_n. Specifically, defining $f_n : \mathbf{A} \mapsto \mathrm{tr}\left(\mathbb{E}\left[Z_n Z_n^\top\right]\mathbf{A}\right)$, we get,

$$|\mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right] - \mathbb{E}\left[Z_n^\top A_n Z_n\right]| = |f_n(A_n) - \mathbb{E}\left[f_n(A_n)\right]| \ .$$

Since f_n is $\|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}}$-Lipschitz, and A_n is κ_n-Lipschitz concentrated, thus from Lemma 5 we get,

$$|f_n(A_n) - \mathbb{E}\left[f_n(A_n)\right]| \prec_s \kappa_n\|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}} \ .$$

The first statement of the corollary follows from the additivity property of stochastic domination (Lemma 4). For the second statement, use $(a+b)^k \leqslant 2^{k-1}(a^k+b^k)$ and take the expectation in (26), to get,

$$\mathbb{E}\left[\left|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n\right]\right|^k\right] \leqslant 2^{k-1}\Bigg(\mathbb{E}\left[\mathbb{E}\left[\left|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]\right|^k \mid A_n\right]\right] + \mathbb{E}\left[\left|f_n(A_n) - \mathbb{E}\left[f_n(A_n)\right]\right|^k\right]\Bigg) .$$

The previous function f_n being $\|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}}$-Lipschitz, we get from the Lipschitz concentration assumption on A_n, as well as the well-known moment bound on sub-Gaussian random variables that:

$$\mathbb{E}\left[\left|f_n(A_n) - \mathbb{E}\left[f_n(A_n)\right]\right|^k\right] \leqslant \mathrm{c}_k \kappa_n^k \|\mathbb{E}\left[Z_n Z_n^\top\right]\|_{\mathrm{F}}^k .$$

Moreover, applying the moment bound of Lemma 6 conditionally on A_n yields the almost sure bound,

$$\mathbb{E}\left[\left|Z_n^\top A_n Z_n - \mathbb{E}\left[Z_n^\top A_n Z_n \mid A_n\right]\right|^k \mid A_n\right] \leqslant \mathrm{c}_k \sigma_n^{2k} \|A_n\|_{\mathrm{F}}^k ,$$

for some large enough universal constants $(\mathrm{c}_k)_{k\in\mathbb{N}} > 0$. The second statement follows. □

C.3 Concentration through a non Lipschitz mapping

This final subsection is dedicated to showing that random variables satisfying Definition 2 preserve satisfying concentration properties when composed with a non Lipschitz mapping.

More precisely, we will consider a differentiable map $f : \mathbb{R}^k \to \mathbb{R}$, and notice that on any compact set $\mathsf{K} \subset \mathbb{R}^k$, the function $f_{|\mathsf{K}}$ is Lipschitz. The concentration of $f(Z)$ will then follow from finding a trade-off between the grows of the Lipschitz constant of $f_{|\mathsf{K}}$ and the probability of the event $\{Z \notin \mathsf{K}\}$. We will denote $\mathtt{L}_{\mathsf{K}}^f$ the Lipschitz constant of f on K:

$$\mathtt{L}_{\mathsf{K}}^f := \sup_{(x,y)\in\mathsf{K}^2} \{|f(x) - f(y)|/\|x-y\|_2\} . \tag{27}$$

In addition, if $\mathtt{L}_{\mathsf{K}}^f < +\infty$, consider the Moreau envelop g_{K}^f of and the Moreau residual δ_{K}^f defined as,

$$g_{\mathsf{K}}^f(z) = \inf_{x\in\mathsf{K}} \{f(x) + \mathtt{L}_{\mathsf{K}}^f \|x-z\|_2\} , \quad \text{and} \quad \delta_{\mathsf{K}}^f(Z) := \mathbb{E}\big[\,|f(Z) - g_{\mathsf{K}}^f(Z)|\,\big] . \tag{28}$$

If $\mathtt{L}_{\mathsf{K}}^f = +\infty$, we use the convention $\delta_{\mathsf{K}}^f = +\infty$.

Lemma 7. *Let $f : \mathbb{R}^k \to \mathbb{C}$ and $\mathsf{K} \subset \mathbb{R}^k$ be a compact set. Assume that $\mathtt{L}_{\mathsf{K}}^f < +\infty$. Then, g_{K}^f defined in (28) is a Lipschitz extension of f restricted to K, i.e., it is a Lipschitz function from $\mathbb{R}^k$ to $\mathbb{R}$ such that for any $x \in \mathsf{K}$, $f(x) = g_{\mathsf{K}}^f(x)$. Furthermore, for any random variable $Z \in \mathbb{R}^k$, and any $z_0 \in \mathsf{K}$, it holds,*

$$\delta_{\mathsf{K}}^f(Z) \leqslant \left(\mathbb{E}\left[|f(Z)|^2\right]^{1/2} + |f(z_0)| + \mathtt{L}_{\mathsf{K}}^f \mathbb{E}\left[\|Z - z_0\|_2^2\right]^{1/2}\right) \mathbb{P}\left(Z \notin \mathsf{K}\right)^{1/2}$$

Proof. Fix $z_1, z_2 \in \mathbb{R}^k$ and any $x \in \mathsf{K}$. By the triangle inequality,

$$f(x) + \mathtt{L}_{\mathsf{K}}^f \|z_1 - x\| \;\leqslant\; f(x) + \mathtt{L}_{\mathsf{K}}^f \big(\|z_2 - x\| + \|z_1 - z_2\|\big) \;=\; \big(f(x) + \mathtt{L}_{\mathsf{K}}^f \|z_2 - x\|\big) + \mathtt{L}_{\mathsf{K}}^f \|z_1 - z_2\|.$$

Taking the infimum over $x \in \mathsf{K}$ yields

$$g_{\mathsf{K}}^f(z_1) \;\leqslant\; g_{\mathsf{K}}^f(z_2) + \mathtt{L}_{\mathsf{K}}^f \|z_1 - z_2\|.$$

Exchanging the roles of z_1 and z_2 gives

$$|g_{\mathsf{K}}^f(z_1) - g_{\mathsf{K}}^f(z_2)| \;\leqslant\; \mathtt{L}_{\mathsf{K}}^f \|z_1 - z_2\|,$$

so g_{K}^f is $\mathtt{L}_{\mathsf{K}}^f$-Lipschitz on $\mathbb{R}^k$. We now show that g_{K}^f coincides with f on K. Let $z \in \mathsf{K}$, since $f_{|\mathsf{K}}$ is $\mathtt{L}_{\mathsf{K}}^f$-Lipschitz, we have for any $x \in \mathsf{K}$,

$$f(z) \leqslant f(x) + \mathtt{L}_{\mathsf{K}}^f \|z - x\|_2 .$$

taking the infimum over $x \in \mathsf{K}$ yields $f(z) \leqslant g^f_{\mathsf{K}}(z)$. Furthermore, taking $x = z$ in the definition of $g^f_{\mathsf{K}}(z)$ gives $g^f_{\mathsf{K}}(z) \leqslant f(z)$, proving the equality. Finally, let $Z \in \mathbb{R}^k$ to be any random vector, without loss of generality (and up to replacing K by $\mathsf{K} - z_0$ and $f(\cdot)$ by $f(\cdot + z_0)$), we assume $z_0 = 0$. Leveraging that $f - g^f_{\mathsf{K}}$ is zero on K, we write,

$$\delta^f_{\mathsf{K}}(Z) = \mathbb{E}\left[|f(Z) - g^f_{\mathsf{K}}(Z)|\mathbb{1}_{\mathsf{K}^c}(Z)\right] \leqslant \mathbb{E}\left[|f(Z)|\mathbb{1}_{\mathsf{K}^c}(Z)\right] + \mathbb{E}\left[|g^f_{\mathsf{K}}(Z)|\mathbb{1}_{\mathsf{K}^c}(Z)\right]$$

We now bound the two terms in the right-hand side. For the first term, we have by the Cauchy-Schwarz inequality,

$$\mathbb{E}\left[|f(Z)|\mathbb{1}_{\mathsf{K}^c}(Z)\right] \leqslant \mathbb{E}\left[|f(Z)|^2\right]^{1/2} \mathbb{P}(Z \notin \mathsf{K})^{1/2} \, .$$

The second term can be bounded thanks to the Lipschitz property of g^f_{K},

$$\begin{aligned}\mathbb{E}\left[|g^f_{\mathsf{K}}(Z)|\mathbb{1}_{\mathsf{K}}(Z)\right] &\leqslant |g^f_{\mathsf{K}}(0)|\mathbb{P}(Z \notin \mathsf{K}) + \mathtt{L}^f_{\mathsf{K}}\mathbb{E}\left[\|Z\|_2 \mathbb{1}_{\mathsf{K}^c}(Z)\right] \\ &\leqslant |f(0)|\mathbb{P}(Z \notin \mathsf{K}) + \mathtt{L}^f_{\mathsf{K}}\mathbb{E}\left[\|Z\|_2^2\right]^{1/2} \mathbb{P}(Z \notin \mathsf{K})^{1/2} \, ,\end{aligned}$$

where the last inequality followed from the Cauchy-Schwarz inequality, as well as the fact that $0 \in \mathsf{K}$. The claimed upper bound follows. □

Lemma 8 (Concentration through locally Lipschitz maps, all $t > 0$)**.** *Let $Z \in \mathbb{R}^k$ be zero-mean and κ–Lipschitz concentrated random vector, and a compact set $\mathsf{K} \subset \mathbb{R}^k$. We have for every $t > 0$,*

$$\mathbb{P}\Big(\, |f(Z) - \mathbb{E}f(Z)| \geqslant t \,\Big) \;\leqslant\; 2\exp\Big(-\frac{t^2}{8\,\kappa^2(\mathtt{L}^f_{\mathsf{K}})^2}\Big) \;+\; \frac{4\,\delta^f_{\mathsf{K}}(Z)}{t}.$$

Proof. The case $\mathtt{L}^f_{\mathsf{K}} = +\infty$ is trivial. We now suppose that $\mathtt{L}^f_{\mathsf{K}} < +\infty$.

Define the function $\Delta^f_{\mathsf{K}} := f - g^f_{\mathsf{K}}$, for any $t > 0$, it holds

$$\mathbb{P}\big(|f(Z) - \mathbb{E}f(Z)| \geqslant t\big) \;\leqslant \mathbb{P}\Big(|g^f_{\mathsf{K}}(Z) - \mathbb{E}g^f_{\mathsf{K}}(Z)| \geqslant t/2\Big) \;+\; \mathbb{P}\Big(|\Delta^f_{\mathsf{K}}(Z) - \mathbb{E}\Delta^f_{\mathsf{K}}(Z)| \geqslant t/2\Big).$$

We now bounds the two terms in the right-hand side. Since g^f_{K} is $\mathtt{L}^f_{\mathsf{K}}$–Lipschitz, and Z is supposed to be κ–Lipschitz concentrated, we get

$$\mathbb{P}\Big(|g^f_{\mathsf{K}}(Z) - \mathbb{E}g^f_{\mathsf{K}}(Z)| \geqslant t/2\Big) \leqslant 2\exp\{-t^2/(8\,\kappa^2(\mathtt{L}^f_{\mathsf{K}})^2)\} \, .$$

For the second term by the triangle inequality and Markov's inequality, we get

$$\mathbb{P}\Big(|\Delta^f_{\mathsf{K}}(Z) - \mathbb{E}\Delta^f_{\mathsf{K}}(Z)| \geqslant t/2\Big) \leqslant \frac{2\,\mathbb{E}|\Delta^f_{\mathsf{K}}(Z) - \mathbb{E}\Delta^f_{\mathsf{K}}(Z)|}{t} = \frac{4\,\delta^f_{\mathsf{K}}}{t} \, .$$

□

Combining the results of Lemma 7 and Lemma 8, one can see that the concentration of $f(Z)$ can be derived from $\mathtt{L}^{\mathsf{K}}_f$ and $\mathbb{P}(Z \notin \mathsf{K})$ being both small, in particular, it holds,

Proposition 3 (Stochastic domination for non-Lipschitz transforms of a Lipschitz concentrated random matrix)**.** *Let $(Z_n)_{n\in\mathbb{N}}$ be a sequence of random matrices, with for all $n \in \mathbb{N}$, $Z_n \in \mathbb{R}^{d_n \times n}$ being zero-mean and κ_n-Lipschitz concentrated. Consider the compact set of $\mathbb{R}^{d_n\times n}$ given by,*

$$\mathsf{K}_n = \Big\{ \boldsymbol{Z} \in \mathbb{R}^{d_n \times n} : \|\boldsymbol{Z}\|_{\mathrm{op}} \leqslant 2\mathrm{a}_1\kappa_n \left(\sqrt{n} + \sqrt{d_n}\right) \Big\} \, ,$$

where $\mathrm{a}_1 > 0$ is defined in (20). In addition, let $f_n : \mathbb{R}^{d_n\times n} \to \mathbb{C}$ be a function holomorphic in its second argument. For $\zeta \in \mathbb{C}$, we denote by $\mathtt{L}_n(\zeta_n)$ the Lipschitz constant of $f_n(\cdot, \zeta_n)$ on K_n defined in (27). For some fixed $\zeta \in \mathbb{C}$ and $r_n > 0$, we set,

$$\mathtt{L}^\star_n(\zeta, r_n) = \sup_{\zeta' \in \mathrm{S}(\zeta, r_n)} \mathtt{L}_n(\zeta') \, , \quad M^{(2)}_n(\zeta) = \mathbb{E}\left[\left|f_n(Z_n, \zeta)\right|^2\right]^{1/2} \, , \quad M^{(2)}_n(\zeta, r_n) = \sup_{\zeta' \in \mathrm{S}(\zeta, r_n)} M^{(2)}_n(\zeta') \, ,$$

In addition, we assume that:

1. *The shapes of the matrices satisfy,*

$$\sup_{n\in\mathbb{N}} \frac{d_n}{n} < +\infty \, .$$

2. *For any $n \in \mathbb{N}$ and $\zeta \in \mathbb{C}$ and $r \geqslant 0$, there exists $\mathtt{D}_n(\zeta, r)$ such that*

$$\sup_{\zeta' \in \mathrm{S}(\zeta,r)} \left| \frac{\partial f_n}{\partial \zeta'}(Z_n, \zeta') - \mathbb{E}\Big[\frac{\partial f_n}{\partial \zeta'}(Z_n, \zeta')\Big] \right| \prec_s \mathtt{D}_n(\zeta, r) \, ,$$

which satisfies for some $\beta_2 > 0$ (independent of ζ and r and n),

$$\sup_{n\in\mathbb{N}} \frac{\mathtt{D}_n(\zeta, r)}{\kappa_n \, \mathtt{L}_n^\star(\zeta, r) n^{\beta_2}} < +\infty \, .$$

Then, for some constants $\mathrm{c} > 0$, it holds for all $z \in \mathbb{C}$ and $r > 0$,

$$\sup_{\zeta' \in \mathrm{S}(\zeta, r_n)} \Big| f_n\big(Z_n, \zeta'\big) - \mathbb{E}\left[f_n\big(Z_n, \zeta'\big)\right] \Big| \prec_s \kappa_n \mathtt{L}_n^\star(\zeta, r)(1 + r) + \Big(M_n^{(2)}(\zeta, r) + \sup_{\mathrm{S}(\zeta, r_n)} |f_n(\mathbf{0}, \zeta')| \Big) \mathrm{e}^{-\mathrm{b}n}.$$

Proof. We first control the probability $\mathbb{P}(Z_n \in \mathsf{K}_n)$. To this end, we leverage that Z_n is centered and has independant and κ_n-Lipschitz concentrated columns, hence by Lemma 2, we have for some $\mathrm{c} > 0$,

$$\mathbb{P}\left(\|Z_n\|_{\mathrm{op}} > 2\mathrm{a}_1 \kappa_n (\sqrt{n} + \sqrt{d_n}) \right) \leqslant 2e^{-\mathrm{c}n} \, .$$

where we used assumption 1 to simplify the inequality. Therefore, we get,

$$\mathbb{P}(Z_n \in \mathsf{K}_n) \leqslant 2\mathrm{e}^{-\mathrm{c}n} \, . \tag{29}$$

The proof of the proposition relies on the similar idea as the one used in Lemma 8, throughout the proof, we fix $\zeta \in \mathbb{C}$ and $r > 0$, and for $\zeta' \in \mathrm{S}(\zeta, r)$, we define

$$g_n(\cdot, \zeta') = g_{\mathsf{K}_n}^{f_n(\cdot, \zeta')}, \quad \text{and} \quad \Delta_n = f_n - g_n \, , \, ,$$

where $g_{\mathsf{K}_n}^{f_n(\cdot,\zeta')}$ was introduced in (28). Using the triangle inequality and denoting $\bar{f}_n = f_n(Z_n, \zeta') - \mathbb{E}\left[f_n(Z_n, \zeta')\right]$, for any $\zeta' \in \mathrm{S}(\zeta, r)$

$$\left|\bar{f}_n\big(Z_n, \zeta'\big)\right| \leqslant \left| g_n\big(Z_n, \zeta'\big) - \mathbb{E}\left[g_n\big(Z_n, \zeta'\big)\right] \right| + \left|\Delta_n\big(Z_n, \zeta'\big)\right| + \mathbb{E}\left[\left|\Delta_n\big(Z_n, \zeta'\big)\right|\right] \tag{30}$$

and we recall from Lemma 7 that $g_n(\cdot, \zeta')$ is $\mathtt{L}_n(\zeta')$-Lipschitz. Since $\Delta_n(\mathbf{Z}_n, \zeta') = 0$ by definition, for any $\mathbf{Z}_n \in \mathsf{K}_n$, relying on (29), we have for any fixed $\delta > 0$,

$$\mathbb{P}\left(\left|\Delta_n\big(Z_n, \zeta'\big)\right| > n^\delta \mathbb{E}\left[\left|\Delta_n\big(Z_n, \zeta'\big)\right|\right]\right) \leqslant \mathbb{P}\left(\left|\Delta_n\big(Z_n, \zeta'\big)\right| > 0\right) \leqslant \mathbb{P}\left(Z_n \notin \mathsf{K}_n\right) \leqslant 2\mathrm{e}^{-\mathrm{c}n} \, ,$$

this ensures for any $\zeta' \in \mathrm{S}(\zeta, r)$,

$$\left|\Delta_n\big(Z_n, \zeta'\big)\right| \prec_s \mathbb{E}\left[\left|\Delta_n\big(Z_n, \zeta'\big)\right|\right] \, . \tag{31}$$

Furthermore, from Lemma 7, and because $\mathbf{0} \in \mathsf{K}_n$ we have for any $\zeta' \in \mathrm{S}(\zeta, r)$

$$\mathbb{E}\left[\left|\Delta_n\big(Z_n, \zeta'\big)\right|\right] \leqslant \Big(M_n^{(2)}(\zeta') + |f_n(\mathbf{0}, \zeta')| + \mathtt{L}_n(\zeta') \|Z_n\|_{\mathrm{F}} \Big) \, \mathbb{P}\left(Z_n \notin \mathsf{K}_n\right)^{1/2}$$

Using our assumptions on the Z_n's, we have,

$$\mathbb{E}\left[\|Z_n\|_{\mathrm{F}}^2\right] = n \mathrm{tr}\left(\mathbb{E}\left[\Sigma_n\right]\right) \, ,$$

where $\Sigma_n = n^{-1}\mathbb{E}\left[Z_n Z_n^\top\right] \in \mathbb{R}^{d_n \times d_n}$ is the covariance matrix of the columns of Z_n. From the concentration assumption, as well as Lemma 1 we have $\mathrm{tr}(\Sigma_n) \leqslant d_n \|\Sigma_n\|_{\mathrm{op}} \leqslant d_n \kappa_n^2$ a.s., therefore, by condition 1, for some $\mathrm{c} > 0$, it holds

$$\mathbb{E}\left[\|Z_n\|_{\mathrm{F}}^2\right] \leqslant n d_n \kappa_n^2 \leqslant \mathrm{c}^2 n^2 \kappa_n^2 \, .$$

Hence, it holds for any $\zeta' \in \mathrm{S}(\zeta, r)$

$$\mathbb{E}\left[\left|\Delta_n(Z_n, \zeta')\right|\right] \leqslant 2\left(M_n^{(2)}(\zeta') + |f_n(\mathbf{0}, \zeta')| + \mathrm{cL}_n(\zeta') n\kappa_n\right) \mathrm{e}^{-\mathrm{b}n/2} \tag{32}$$

where we used (29). By additivity of $\prec_s$ Lemma 4, (30)-(31)-(32) imply that for all $\delta, D > 0$, there exists $N(\delta, D)$ such that for all $n \geqslant N(\delta, D)$ and and $\zeta' \in \mathrm{S}(\zeta, r)$, and with probability at least $1 - n^{-D}$,

$$\left|\bar{f}_n(Z_n, \zeta')\right| \leqslant n^{\delta}\left|g_n(Z_n, \zeta') - \mathbb{E}\left[g_n(Z_n, \zeta')\right]\right| \quad + 2n^{\delta}\left(M_n^{(2)}(\zeta') + |f_n(\mathbf{0}, \zeta')| + \mathrm{cL}_n(\zeta') n\kappa_n\right) \mathrm{e}^{-\mathrm{b}n/2} .$$

We further upper bound the term involving g_n, by leveraging the Lipschitz property of $g_n(\cdot, \zeta')$ and the Lipschitz-concentration of Z_n, we have for all $t \geqslant 0$,

$$\mathbb{P}\left(\left|g_n(Z_n, \zeta') - \mathbb{E}\left[g_n(Z_n, \zeta')\right]\right| \geqslant t\right) \leqslant 2\exp\left(\frac{-t^2}{2\kappa_n^2 \mathrm{L}_n(\zeta')^2}\right) ,$$

thus taking $t = n^{\delta}\kappa_n \mathrm{L}_n(\zeta')$, we get,

$$\mathbb{P}\left(\left|g_n(Z_n, \zeta') - \mathbb{E}\left[g_n(Z_n, \zeta')\right]\right| \geqslant n^{\delta}\kappa_n \mathrm{L}_n(\zeta')\right) \leqslant 2\exp\left(\frac{-n^{2\delta}}{2}\right)$$

and it holds with probability at least $1 - 2\exp(-n^{2\delta}/2) - n^{-D'}$ that

$$\left|\bar{f}_n(Z_n, \zeta')\right| \leqslant n^{\delta}\left\{\kappa_n \mathrm{L}_n(\zeta') + \left(M_n^{(2)}(\zeta') + |f_n(\mathbf{0}, \zeta')| + \mathrm{cL}_n(\zeta') n\kappa_n\right) \mathrm{e}^{-\mathrm{b}n/2}\right\} . \tag{33}$$

This concludes the proof in the case $r_n = 0$, for the general case, we use a ε-net technique to derive a control uniform over $\mathrm{S}(\zeta, r_n)$. Consider for $\varepsilon > 0$, the set

$$\mathsf{N}_{\varepsilon}(\zeta, r) := \left\{\zeta + r\mathrm{e}^{2\mathrm{i}\pi\ell/k} : 0 \leqslant \ell \leqslant \lceil 1/\varepsilon \rceil - 1\right\} ,$$

and for any $\zeta' \in \mathbb{C}$, $\mathrm{proj}(\zeta') \in \arg\min_{\tilde{\zeta} \in \mathsf{N}_{\varepsilon}(\zeta, r)} \left|\zeta' - \tilde{\zeta}\right|$. Therefore, considering the parameterization $\theta \in [0, 2\pi] \mapsto \zeta'(\theta) = z + r\mathrm{e}^{i\theta}$, using a Taylor expansion and condition 2, we have for any $\zeta' \in \mathrm{S}(\zeta, r)$,

$$\sup_{\zeta' \in \mathrm{S}(\zeta, r)} \left|\bar{f}_n(Z_n, \zeta') - \bar{f}_n(Z_n, \mathrm{proj}(\zeta'))\right| \prec_s r\mathtt{D}_n(\zeta, r)\varepsilon ,$$

It yields

$$\begin{aligned}\sup_{\zeta' \in \mathrm{S}(\zeta, r)} &\left|f_n(Z_n, \zeta') - \mathbb{E}\left[f_n(Z_n, \zeta')\right]\right| \\ &\prec_s \max_{\zeta' \in \mathsf{N}_{\varepsilon}} \left|f_n(Z_n, \zeta') - \mathbb{E}\left[f_n(Z_n, \zeta')\right]\right| + r\,\mathtt{D}_n(\zeta, r)\,\varepsilon .\end{aligned} \tag{34}$$

By a union bound, using (33), and setting

$$t_n = \kappa_n \mathrm{L}_n^{\star}(\zeta, r_n) + \left(M_n^{(2)}(\zeta, r_n) + |f_n(\mathbf{0}, \zeta')| + \mathrm{cL}_n^{\star}(\zeta, r_n) n\kappa_n\right) \mathrm{e}^{-\mathrm{b}n/2} ,$$

we get for all $\delta > 0$, and $D > 0$, there exists $N(\delta, D)$ such that for any $n \geqslant N(\delta, D)$,

$$\begin{aligned}\mathbb{P}\left(\max_{\zeta' \in \mathsf{N}_{\varepsilon}} \left|f_n(Z_n, \zeta') - \mathbb{E}\left[f_n(Z_n, \zeta')\right]\right| \geqslant n^{\delta} t_n\right) &\leqslant \sum_{\zeta' \in \mathsf{N}_{\varepsilon}} \mathbb{P}\left(\left|f_n(Z_n, \zeta') - \mathbb{E}\left[f_n(Z_n, \zeta')\right]\right| \geqslant n^{\delta} t_n\right) \\ &\leqslant |\mathsf{N}_{\varepsilon}|\left\{2\exp\left(-n^{2\delta}/2\right) + n^{-D}\right\} .\end{aligned}$$

Finally, we choose $\varepsilon = \kappa_n \mathrm{L}_n^{\star}(\zeta, r_n)/\mathtt{D}_n(\zeta, r_n)$, so that (34) implies that for all $\delta > 0$, and $D > 0$, there exists $N(\delta, D)$ such that for any $n \geqslant N(\delta, D)$,

$$\mathbb{P}\left(\sup_{\zeta' \in \mathrm{S}(\zeta, r_n)} \left|f_n(Z_n, \zeta') - \mathbb{E} f_n(Z_n, \zeta')\right| \geqslant n^{\delta} t_n + r_n\kappa_n\right) \leqslant \frac{r\mathtt{D}_n(z, r)}{\kappa_n \mathrm{L}_n^{\star}(z, r)}\left\{2k\exp\left(-n^{2\delta}/2\right) + n^{-D}\right\}$$

Condition 2 implies that for any $\delta, D' > 0$, there exists $N(\delta, D')$ such that setting $D = D' + 1 + \beta_2$, it holds

$$\frac{r\mathtt{D}_n(z,r)}{\mathtt{L}_n^\star(\zeta,r_n)\kappa_n}\left\{2k\exp\left(-n^{2\delta}/2\right)+n^{-D'}\right\}\leqslant n^{-D'}\ .$$

By the definition of stochastic domination, this shows, for any $\delta, D' > 0$, there exists $N(\delta, D')$ such that setting $D = D' + 1 + \beta_2$, it holds

$$\sup_{\zeta'\in\mathrm{S}(\zeta,r)}\left|f_n(Z_n,\zeta')-\mathbb{E}f_n(Z_n,\zeta')\right|\prec_s t_n + r_n\mathtt{L}_n^\star(\zeta,r_n)\kappa_n\ .$$

Which simplifies into

$$\sup_{\zeta'\in\mathrm{S}(\zeta,r_n)}\left|f_n(Z_n,\zeta')-\mathbb{E}f_n(Z_n,\zeta')\right|\prec_s \kappa_n(1+r_n)\mathtt{L}_n^\star(\zeta,r_n)+\left(M_n^{(2)}(\zeta,r_n)+|f_n(\mathbf{0},\zeta')|\right)\mathrm{e}^{-\mathrm{b}n/2}\ .$$

and concludes the proof. □

D Concentration of $\mathcal{G}(\alpha,\lambda)$ in the proportional regime

Building upon the preceding result, we are able to establish sharp concentration inequalities for the augmented resolvent matrix and related quantities. Specifically, we demonstrate that this result allows to obtain high-probability bounds for a range of transformations of $\hat{\theta}_{\alpha,\lambda}$, ultimately yielding the concentration of $\mathcal{G}(\alpha,\lambda)$ as defined in (4). We will consider quantities of the form,

$$\begin{aligned}\mathrm{R}_{\alpha,\lambda}(\zeta) &= \left((1-\alpha)\mathrm{C}+\alpha\mathrm{C}'+\alpha\Lambda(\mathtt{X},\mathtt{Y})+\lambda\mathrm{I}_d+\zeta\Sigma\right)^{-1}\ ,\\ \mathrm{H}_\alpha &= \left((1-\alpha)\mathrm{H}+\alpha\mathrm{H}'+\alpha\Omega(\mathtt{X},\mathtt{Y})\right)\\ \text{and}\quad \xi_{\alpha,\lambda}(\zeta) &= \mathrm{H}_\alpha^\top\mathrm{R}_{\alpha,\lambda}(\zeta)\mathrm{H}_\alpha\ ,\end{aligned}\tag{35}$$

for an $\zeta\in\mathsf{U}_\lambda=\{\zeta\in\mathbb{C}:\Re(\zeta)>-\lambda/\mathtt{L}_\varphi^2\}$. Note that $\zeta\in\mathsf{U}_\lambda$, Lemma 1 and **H**1 ensure that $\min\{\Re(\lambda+\rho):\rho\in\mathrm{Sp}(\zeta\Sigma)\}>0$, thus garanteeing that $\xi_{\alpha,\lambda}(\zeta)$ is well-defined. With these notations, $\mathcal{G}(\alpha,\lambda)$ in (14) rewrites as

$$\mathcal{G}(\alpha,\lambda)=\theta_\star^\top\Sigma_{\star\star}\theta_\star+\sigma^2-2\theta_\star^\top\Sigma_\star\hat{\theta}_{\alpha,\lambda}-\partial_\zeta\xi_{\alpha,\lambda}(0)\ ,\tag{36}$$

for any $\alpha\in[0,1]$ and $\lambda>0$. Our objective is to show that, with high probability, $\mathcal{G}(\alpha,\lambda)$ remains close to its mean. To this end, we will invoke Proposition 3 on $\theta_\star^\top\Sigma_\star\hat{\theta}_{\alpha,\lambda}$ and $\xi_{\alpha,\lambda}(\zeta)$ respectively. The concentration of the derivative $\partial_\zeta\xi_{\alpha,\lambda}(0)$ then follows from an appropriate differentiation lemma, formulated below in the formalism of stochastic domination.

Lemma 9 (Differentiation under uniform stochastic control)**.** *Let $(U_n)_{n\in\mathbb{N}}$ and $(V_n)_{n\in\mathbb{N}}$ be sequences of random holomorphic functions from $\mathbb{C}$ to $\mathbb{C}$. Fix $\zeta\in\mathbb{C}$ and $r>0$. Suppose there exists $\xi_n(\zeta,r)>0$ such that*

$$\sup_{\zeta'\in\mathrm{S}(\zeta,r)}\left|U_n(\zeta')-V_n(\zeta')\right|\ \prec_s\ \xi_n(\zeta,r)\ .\tag{37}$$

Then, the derivatives at ζ satisfy

$$\left|\frac{\partial U_n}{\partial\zeta}(\zeta)-\frac{\partial V_n}{\partial\zeta}(\zeta)\right|\ \prec_s\ \frac{\xi_n(\zeta,r)}{r}\ .\tag{38}$$

Proof. By Cauchy's integral formula for the derivative of a holomorphic function,

$$\frac{\partial}{\partial\zeta}(U_n-V_n)(\zeta)=\frac{1}{2\pi\mathrm{i}}\oint_{\zeta'\in\mathrm{S}(\zeta,r)}\frac{U_n(\zeta')-V_n(\zeta')}{(\zeta'-\zeta)^2}\,\mathrm{d}\zeta'\ .$$

Taking the module and noting that the maximum of $|U_n(\zeta')-V_n(\zeta')|$ over the circle $\mathrm{S}(\zeta,r)$ controls the integral, we obtain

$$\left|\frac{\partial}{\partial\zeta}(U_n-V_n)(\zeta)\right|\leqslant\frac{1}{r}\sup_{\zeta'\in\mathrm{S}(\zeta,r)}|U_n(\zeta')-V_n(\zeta')|\ .$$

By the stochastic domination assumption (37), we obtain the desired bound (38). □

In order to apply this result to the sequence $\partial_\zeta \xi_{\alpha,\lambda_n}(0)$, we first prove the following result:

Proposition 4. *For all $n \in \mathbb{N}$, let $\lambda_n > 0$, $0 \leqslant r_n < \lambda_n / \mathrm{L}_\varphi^2$ and $\mathbf{a}_n \in \mathbb{R}^p$. Assume that **H2**, **H1** and **H4** hold. Then, we have,*

$$\left| \mathbf{a}_n^\top \hat{\theta}_{\alpha,\lambda_n} - \mathbb{E}\left[\mathbf{a}_n^\top \hat{\theta}_{\alpha,\lambda_n} \right] \right| \prec_s \frac{\|\mathbf{a}_n\|_2}{\lambda_n^{3/2}\sqrt{n}} \left(1 + \sqrt{\lambda_n} \right) ,$$

and,

$$\sup_{t\in[0,2\pi]} \left| \xi_{\alpha,\lambda_n}(r_n \mathrm{e}^{\mathrm{i}t}) - \mathbb{E}\left[\xi_{\alpha,\lambda_n}(r_n \mathrm{e}^{\mathrm{i}t}) \right] \right| \prec_s \frac{1 + r_n}{(\lambda_n - r_n \mathrm{L}_\varphi^2)^{3/2}\sqrt{n}} \left(1 + \sqrt{\lambda_n - r_n \mathrm{L}_\varphi^2} \right) .$$

Before proving Proposition 4, we first introduce the following notations for the smallest/largest real parts of eigenvalues associated to any complex matrix $\mathbf{D} \in \mathbb{C}^{d\times d}$:

$$\lambda_{\min}^{\Re}(\mathbf{D}) := \min_{\lambda \in \mathrm{Sp}(\mathbf{D})} \Re(\lambda) , \quad \text{and} \quad \lambda_{\max}^{\Re}(\mathbf{D}) := \max_{\lambda \in \mathrm{Sp}(\mathbf{D})} \Re(\lambda) ,$$

which are used in the following lemma on the Lipschitz property of the resolvent matrix with complex regularizations:

Lemma 10. *For any $n, p \in \mathbb{N}$, let $\mathbf{D} \in \mathbb{C}^{p\times p}$ be a complex matrix, and $\mathbf{X}_1, \mathbf{X}_2 \in \mathbb{R}^{p\times n}$ be two matrices. Assume that $\mathrm{Sp}(\mathbf{D}) \subset \{\zeta \in \mathbb{C} : \Re(\zeta) > 0\}$, then we have,*

$$\left\| \left(n^{-1}\mathbf{X}_1\mathbf{X}_1^\top + \mathbf{D} \right)^{-1} - \left(n^{-1}\mathbf{X}_2\mathbf{X}_2^\top + \mathbf{D} \right)^{-1} \right\|_{\mathrm{F}} \leqslant \frac{1}{\sqrt{n}\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}} \|\mathbf{X}_1 - \mathbf{X}_2\|_{\mathrm{F}} ,$$

and

$$\left\| \left(n^{-1}\mathbf{X}_1\mathbf{X}_1^\top + \mathbf{D} \right)^{-1} - \left(n^{-1}\mathbf{X}_2\mathbf{X}_2^\top + \mathbf{D} \right)^{-1} \right\|_{\mathrm{op}} \leqslant \frac{1}{\sqrt{n}\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}} \|\mathbf{X}_1 - \mathbf{X}_2\|_{\mathrm{op}} .$$

Proof. Write $\|\cdot\|$ to represent either $\|\cdot\|_{\mathrm{op}}$ or $\|\cdot\|_{\mathrm{F}}$. We first use the resolvent identity to write,

$$\begin{aligned}
&\left\| \left(n^{-1}\mathbf{X}_1\mathbf{X}_1^\top + \mathbf{D} \right)^{-1} - \left(n^{-1}\mathbf{X}_2\mathbf{X}_2^\top + \mathbf{D} \right)^{-1} \right\| \\
=& p^{-1} \left\| \left(n^{-1}\mathbf{X}_1\mathbf{X}_1^\top + \mathbf{D} \right)^{-1} \left\{ \mathbf{X}_1\mathbf{X}_1^\top - \mathbf{X}_2\mathbf{X}_2^\top \right\} \left(n^{-1}\mathbf{X}_2\mathbf{X}_2^\top + \mathbf{D} \right)^{-1} \right\| \\
=& p^{-1} \left\| \left(n^{-1}\mathbf{X}_1\mathbf{X}_1^\top + \mathbf{D} \right)^{-1} \left\{ \mathbf{X}_1(\mathbf{X}_1 - \mathbf{X}_2)^\top + (\mathbf{X}_1 - \mathbf{X}_2)\mathbf{X}_2^\top \right\} \left(n^{-1}\mathbf{X}_2\mathbf{X}_2^\top + \mathbf{D} \right)^{-1} \right\| \\
\leqslant& \frac{\|(\mathbf{C}_2 + \mathbf{D})^{-1}\|_{\mathrm{op}} \|\left(\mathbf{C}_1 + \mathbf{D}\right)^{-1}\mathbf{X}_1\|_{\mathrm{op}} + \|(\mathbf{C}_1 + \mathbf{D})^{-1}\|_{\mathrm{op}} \|\left(\mathbf{C}_2 + \mathbf{D}\right)^{-1}\mathbf{X}_2\|_{\mathrm{op}}}{n} \|\mathbf{X}_1 - \mathbf{X}_2\| ,
\end{aligned} \tag{39}$$

where we have introduced the notation $\mathbf{C}_i = n^{-1}\mathbf{X}_i\mathbf{X}_i^\top$ for $i \in \{1, 2\}$. Furthermore, one can check that $\|(\mathbf{C}_i + \mathbf{D})^{-1}\|_{\mathrm{op}} = \max_{\lambda\in\mathrm{Sp}(\mathbf{C}_i+\mathbf{D})} |\lambda|^{-1} \leqslant 1/\lambda_{\min}^{\Re}(\mathbf{D})$. Additionally, we bound the remaining operator norm, for both $i \in \{1, 2\}$, it holds,

$$\begin{aligned}
\left\| \left(\mathbf{C}_i + \mathbf{D}\right)^{-1}\mathbf{X}_i^\top \right\|_{\mathrm{op}}^2 &= n\, \lambda_{\max}\Big(\left(\mathbf{C}_i + \mathbf{D}\right)^{-1} \mathbf{C}_i \left(\left(\mathbf{C}_i + \mathbf{D}\right)^{-1} \right)^\dagger \Big) \\
&= n \sup_{v\in\mathbb{C}^n\setminus\{0\}} \frac{v^\dagger (\mathbf{C}_i + \mathbf{D})^{-1}\mathbf{C}_i(\mathbf{C}_i + \mathbf{D}^\dagger)^{-1}v}{\|v\|_2^2} \\
&= n \sup_{w\in\mathbb{C}^n\setminus\{0\}} \frac{w^\dagger \mathbf{C}_i w}{\|(\mathbf{C}_i + \mathbf{D}^\dagger)w\|_2^2} .
\end{aligned}$$

where we used the change of variable $w = (\mathbf{C}_i + \mathbf{D}^\dagger)^{-1}v$ in the last line. Now remark that from Cauchy-Schwarz inequality, it holds,

$$\begin{aligned}
\|(\mathbf{C}_i + \mathbf{D}^\dagger)w\|_2 \geqslant \frac{\left|\left\langle (\mathbf{C}_i + \mathbf{D}^\dagger)w, w \right\rangle\right|}{\|w\|_2} &\geqslant \frac{\Re\left(\left\langle (\mathbf{C}_i + \mathbf{D}^\dagger)w, w \right\rangle\right)}{\|w\|_2} \\
&\geqslant \frac{w^\dagger \mathbf{C}_i w + \lambda_{\min}^{\Re}(\mathbf{D})\|w\|_2^2}{\|w\|_2} .
\end{aligned}$$

Hence, defining $c = w^\dagger \mathbf{C}_i w / \|w\|_2^2 \geqslant 0$, we have,

$$\left\| (\mathbf{C}_i + \mathbf{D})^{-1} \mathbf{X}_i^\top \right\|_{\mathrm{op}}^2 \leqslant \sup_{c \geqslant 0} \frac{c}{(c + \lambda_{\min}^{\Re}(\mathbf{D}))^2} \, .$$

To conclude the proof, we simply remark that the function $c \mapsto c/(c + \lambda_{\min}^{\Re}(\mathbf{D}))^2$ is maximized at $c = \lambda_{\min}^{\Re}(\mathbf{D})$, which yields,

$$\left\| (\mathbf{C}_i + \mathbf{D})^{-1} \mathbf{X}_i^\top \right\|_{\mathrm{op}}^2 \leqslant \frac{1}{4\lambda_{\min}^{\Re}(\mathbf{D})} \, .$$

Plugging this result in (39) concludes the proof. □

Proof of Proposition 4. Fir $\mathbf{a}_n \in \mathbb{R}^p$. First note that by assumption U_{λ_n} contains $\{\lambda_n + r_n \mathrm{e}^{\mathrm{i}t} : t \in [0, 2\pi)\}$ for all $n \in \mathbb{N}$, thus $\xi_{\alpha,\lambda_n}(r_n \mathrm{e}^{\mathrm{i}t})$ is well-defined for all $t \in [0, 2\pi]$. The proof heavily relies on Proposition 3, we first exhibit two functions $q : \mathbb{R}^{d \times n} \to \mathbb{R}$ and $s : \mathbb{R}^{d \times n} \times \mathbb{C} \to \mathbb{C}$ such that

$$\mathbf{a}_n^\top \hat{\theta}_{\alpha,\lambda_n} = q(\mathtt{X}, \mathtt{Y}) \, , \quad \text{and} \quad \xi_{\alpha,\lambda_n}(r_n \mathrm{e}^{\mathrm{i}t}) = s(\mathtt{X}, \mathtt{Y}, r_n \mathrm{e}^{\mathrm{i}t}) \, .$$

To this end, we define for any $\mathbf{X} \in \mathbb{R}^{d \times n}$, $\mathbf{Y} \in \mathbb{R}^{d \times n}$, and $\zeta \in \mathbb{C}$,

$$\begin{aligned} r(\mathbf{X}, \mathbf{Y}, \zeta) &= (n^{-1} \varphi(\mathbf{X}) \varphi(\mathbf{X})^\top + n^{-1} \mu_x(\mathbf{X}, \mathbf{Y}) \mu_x(\mathbf{X}, \mathbf{Y})^\top + \alpha \Lambda(\mathbf{X}, \mathbf{Y}) + \lambda_n \mathrm{I}_d + \zeta \Sigma)^{-1} \\ h(\mathbf{X}, \mathbf{Y}, \zeta) &= n^{-1} \varphi(\mathbf{X}) \mathbf{Y}^\top + n^{-1} \mu_x(\mathbf{X}, \mathbf{Y}) \mu_y(\mathbf{X}, \mathbf{Y})^\top + \alpha \Omega(\mathbf{X}, \mathbf{Y}) \, . \end{aligned} \tag{40}$$

where μ_x, μ_y, Λ and Ω were introduced in **H**4. r and h allow to constructions q and s as follows:

$$\begin{aligned} \forall (\mathbf{X}, \mathbf{Y}, \zeta) \in \mathbb{R}^{d \times n} \times \mathbb{R}^{d \times n} \times \mathbb{C}, \quad q(\mathbf{X}, \mathbf{Y}) &= \mathbf{a}_n^\top r(\mathbf{X}, \mathbf{Y}, 0) h(\mathbf{X}, \mathbf{Y}) \, , \\ \text{and} \quad s(\mathbf{X}, \mathbf{Y}, \zeta) &= h(\mathbf{X}, \mathbf{Y})^\top r(\mathbf{X}, \mathbf{Y}, \zeta) h(\mathbf{X}, \mathbf{Y}) \, . \end{aligned}$$

In line with the hypothesis of Proposition 3, we now show that $q(\cdot, \cdot)$ and $s(\cdot, \cdot, \zeta)$ are Lipschitz continuous on the compact set

$$\mathsf{K}_n = \left\{ (\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{d \times n} \times \mathbb{R}^{1 \times n} : \|(\mathbf{X}^\top, \mathbf{Y}^\top)\|_{\mathrm{op}} \leqslant 2\mathrm{a}(\sqrt{n} + \sqrt{d+1}) \right\} \, ,$$

where $\mathrm{a} > 0$ is defined in (20). To this end, let us first derive $\mathtt{L}_{\mathsf{K}_n}^r(\zeta)$ and $\mathtt{L}_{\mathsf{K}_n}^h$, the Lipschitz constants of $r(\cdot, \cdot, \zeta)$ and $h(\cdot, \cdot, \zeta)$ on K_n. Fix any $\zeta \in \mathsf{U}_{\lambda_n}$ and $(\mathbf{X}, \mathbf{Y})$ and $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})$ in K_n, we have from Lemma 10, the Lipschitz property assumption on Λ and μ_x, from **H**4, and the lipschitz property of $\mathbf{M} \mapsto (\mathrm{I}_d + \mathbf{M})^{-1}$ (provided $\Re(\mathrm{Sp}(\mathbf{M})) \subset \mathbb{R}_+$),

$$\begin{aligned} &\left\| r(\mathbf{X}, \mathbf{Y}, \zeta) - r(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}, \zeta) \right\|_{\mathrm{op}} \\ &\leqslant \frac{\sqrt{2}}{\lambda_{\min}^{\Re}(\Lambda(\mathbf{X}, \mathbf{Y}) + \lambda \mathrm{I}_d + \zeta \Sigma)^{3/2} \sqrt{n}} \left\| (\varphi(\mathbf{X}), \mu_x(\mathbf{X}, \mathbf{Y})) - (\varphi(\widetilde{\mathbf{X}}), \mu_x(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}})) \right\|_{\mathrm{op}} \\ &\quad + \frac{1}{\lambda_{\min}^{\Re}(\Lambda(\mathbf{X}, \mathbf{Y}) + \lambda \mathrm{I}_d + \zeta \Sigma)} \left\| \Lambda(\mathbf{X}, \mathbf{Y}) - \Lambda(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \right\|_{\mathrm{op}} \\ &\leqslant \frac{\mathrm{c}(1 + \lambda_{\min}^{\Re}(\Lambda(\mathbf{X}, \mathbf{Y}) + \lambda \mathrm{I}_d + \zeta \Sigma)^{1/2})}{\lambda_{\min}^{\Re}(\Lambda(\mathbf{X}, \mathbf{Y}) + \lambda \mathrm{I}_d + \zeta \Sigma)^{3/2} \sqrt{n}} \left\| (\mathbf{X}, \mathbf{Y}) - (\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \right\|_{\mathrm{op}} \\ &\leqslant \frac{\mathrm{c}(1 + (\lambda_n - |\zeta| \mathtt{L}_\varphi^2)^{1/2})}{(\lambda_n - |\zeta| \mathtt{L}_\varphi^2)^{3/2} \sqrt{n}} \left\| (\mathbf{X}, \mathbf{Y}) - (\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \right\|_{\mathrm{op}} \, , \end{aligned}$$

where we have used $\|\Sigma\|_{\mathrm{op}} \leqslant \mathtt{L}_\varphi^2$, which follows from **H**1 and Lemma 1. It results,

$$\mathtt{L}_{\mathsf{K}_n}^r(\zeta) \leqslant \frac{\mathrm{c}(1 + (\lambda_n - r_n \mathtt{L}_\varphi^2)^{1/2})}{(\lambda_n - r_n \mathtt{L}_\varphi^2)^{3/2} \sqrt{n}} \, . \tag{41}$$

Similarly for h, we use the fact that $(\mathbf{X}, \mathbf{Y}) \in \mathsf{K}_n$ and $(\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}) \in \mathsf{K}_n$ to write,

$$\begin{aligned}\left\|h(\mathbf{X},\mathbf{Y}) - h(\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})\right\|_{\mathrm{op}} \leqslant \frac{\mathrm{c}\left(\sqrt{n}+\sqrt{d}\right)}{n}\Big(&\left\|\mathbf{Y}-\widetilde{\mathbf{Y}}\right\|_{\mathrm{op}} + \left\|\mu_y(\mathbf{X},\mathbf{Y}) - \mu_y(\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})\right\|_{\mathrm{op}} \\ &+ \left\|\mathbf{X}-\widetilde{\mathbf{X}}\right\|_{\mathrm{F}} + \left\|\mu_x(\mathbf{X},\mathbf{Y}) - \mu_x(\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})\right\|_{\mathrm{op}}\Big) \\ &+ \left\|\Omega(\mathbf{X},\mathbf{Y}) - \Omega(\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})\right\|_{\mathrm{op}} \\ \leqslant \frac{\mathrm{c}}{\sqrt{n}}&\left\|(\mathbf{X},\mathbf{Y}) - (\widetilde{\mathbf{X}},\widetilde{\mathbf{Y}})\right\|_{\mathrm{op}} ,\end{aligned}$$

thus,

$$\mathtt{L}^h_{\mathsf{K}_n} \leqslant \frac{\mathrm{c}}{\sqrt{n}} . \tag{42}$$

Combining (41) and (42), as well as using the fact that for any $(\mathbf{X}, \mathbf{Y}) \in \mathsf{K}_n$,

$$\|r(\mathbf{X},\mathbf{Y},\zeta)\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_n - |\zeta|\mathtt{L}^2_\varphi} , \quad \text{and} \quad \|h(\mathbf{X},\mathbf{Y})\|_{\mathrm{op}} \leqslant \mathrm{c} ,$$

we obtain,

$$\mathtt{L}^q_{\mathsf{K}_n} \leqslant \frac{\|\mathbf{a}_n\|_2 \mathtt{L}^h_{\mathsf{K}_n}}{\lambda_n} + \mathrm{c}\|\mathbf{a}_n\|_2 \mathtt{L}^r_{\mathsf{K}_n}(0) \leqslant \frac{\mathrm{c}\|\mathbf{a}_n\|_2(1+\lambda_n^{1/2})}{\lambda_n^{3/2}\sqrt{n}} .$$

and,

$$\mathtt{L}^s_{\mathsf{K}_n}(\zeta) \leqslant \frac{\mathrm{c}(1+(\lambda_n - r_n\mathtt{L}^2_\varphi)^{1/2})}{(\lambda_n - r_n\mathtt{L}^2_\varphi)^{3/2}\sqrt{n}} .$$

We now check that $(\mathtt{X}, \mathtt{Y})$ validates the assumptions of Proposition 3 for the function q and s. First remark that Item 1 of Proposition 3 is guarenteed by **H**2. We thus focus on proving that Item 2 of Proposition 3 holds. Note that q satisfies Item 2 trivially. To prove that s satisfies Item 2, we first note that for any $t \in [0, 2\pi]$,

$$\partial_\zeta s(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t}) = -h(\mathtt{X},\mathtt{Y})^\top r(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t})\Sigma r(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t})h(\mathtt{X},\mathtt{Y}) ,$$

additionally we write,

$$\begin{aligned}\left|\partial_\zeta s(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t}) - \mathbb{E}\left[\partial_\zeta s(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t})\right]\right| &\leqslant \left|\partial_\zeta s(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t})\right| + \mathbb{E}\left[\left|\partial_\zeta s(\mathtt{X},\mathtt{Y},r_n\mathrm{e}^{\mathrm{i}t})\right|\right] \\ &\leqslant \frac{1}{\lambda_n - r_n\mathtt{L}^2_\varphi}\left\|h(\mathtt{X},\mathtt{Y})^\top\right\|_2^2 + \frac{1}{\lambda_n - r_n\mathtt{L}^2_\varphi}\mathbb{E}\left[\left\|h(\mathtt{X},\mathtt{Y})^\top\right\|_2^2\right] \\ &\prec_s \frac{1}{\lambda_n - r_n\mathtt{L}^2_\varphi} + \frac{1}{\lambda_n - r_n\mathtt{L}^2_\varphi}\mathbb{E}\left[\left\|h(\mathtt{X},\mathtt{Y})^\top\right\|_2^2\right]\end{aligned}$$

Furthermore, from the Lipschitz properties of φ, μ_x, μ_y and Ω, we have,

$$\left\|h(\mathtt{X},\mathtt{Y})^\top\right\|_2^2 \leqslant \frac{\mathrm{c}}{n}\mathbb{E}\left[\|(\mathtt{X},\mathtt{Y})\|^2_{\mathrm{op}}\right] \leqslant \mathrm{c}' =: \mathrm{D}_n(r_n\mathrm{e}^{\mathrm{i}t}) ,$$

It results that the ratio $\mathrm{D}_n(r_n\mathrm{e}^{\mathrm{i}t})/\mathtt{L}^q_{\mathsf{K}_n}(r_n\mathrm{e}^{\mathrm{i}t})$ is bounded by a constant, independant of t thus Item 2 of Proposition 3 holds. Having established the correctness of the assumptions of Proposition 3 for q and s, the statement follows by applying Proposition 3. □

Combining previous results, we obtain the following statement, which garentees the concentration of $\mathcal{G}(\alpha, \lambda)$ around its mean:

Theorem 2. *For all $n \geqslant 0$, let $\lambda_n > 0$ and $r_n \geqslant 0$, and assume **H**2, **H**1 and **H**4 hold. Then the following concentration bounds hold:* (4) *satisfies,*

$$\left|\theta_*^\top \Sigma_* \hat{\theta}_{\alpha,\lambda_n} - \mathbb{E}\left[\theta_*^\top \Sigma_* \hat{\theta}_{\alpha,\lambda_n}\right]\right| \prec_s \frac{\|\beta_\star\|_2}{\lambda_n^{3/2}\sqrt{n}}\left(1+\sqrt{\lambda_n}\right) ,$$

$$\sup_{t\in[0,2\pi]} \left|\xi_{\alpha,\lambda_n}(r_n e^{it}) - \mathbb{E}\left[\xi_{\alpha,\lambda_n}(r_n e^{it})\right]\right| \prec_s \frac{1+r_n}{(\lambda_n - r_n \mathtt{L}_\varphi^2)^{3/2}\sqrt{n}} \left(1+\sqrt{\lambda_n - r_n\mathtt{L}_\varphi^2}\right) ,$$

and the generalization error satisfies,

$$|\mathcal{G}(\alpha,\lambda_n) - \mathbb{E}[\mathcal{G}(\alpha,\lambda_n)]| \prec_s \frac{1+\lambda_n^2}{\lambda_n^{5/2}\sqrt{n}} + \frac{\|\beta_\star\|_2(1+\sqrt{\lambda_n})}{\lambda_n^{3/2}\sqrt{n}} .$$

Proof. We leverage (36), as well as the previously derived concentration results for ξ_{α,λ_n} and $\theta_*^\top \Sigma_* \hat{\theta}_{\alpha,\lambda_n}$. Indeed we have, from Proposition 4 that,

$$\left|\theta_*^\top \Sigma_* \hat{\theta}_{\alpha,\lambda_n} - \mathbb{E}\left[\theta_*^\top \Sigma_* \hat{\theta}_{\alpha,\lambda_n}\right]\right| \prec_s \frac{\|\beta_\star\|_2}{\lambda_n^{3/2}\sqrt{n}} \left(1+\sqrt{\lambda_n}\right) , \tag{43}$$

where the term $\|\Sigma_\star\|_{\mathrm{op}} \leqslant \mathtt{L}_\varphi^2$ has been stashed inside the $\prec_s$ notation, by using Lemma 1. Additionally, from Proposition 4 applied with $0 < r_n < \lambda/\mathtt{L}_\varphi^2$, we get,

$$\sup_{t\in[0,2\pi]} \left|\xi_{\alpha,\lambda_n}(r_n e^{it}) - \mathbb{E}\left[\xi_{\alpha,\lambda_n}(r_n e^{it})\right]\right| \prec_s \frac{1+r_n}{(\lambda_n - r_n \mathtt{L}_\varphi^2)^{3/2}\sqrt{n}} \left(1+\sqrt{\lambda_n - r_n\mathtt{L}_\varphi^2}\right) .$$

Combining the previous with Lemma 9, we have, for all $r < \lambda/\mathtt{L}_\varphi^2$,

$$|\partial_\zeta \xi_{\alpha,\lambda_n}(0) - \mathbb{E}\left[\partial_\zeta \xi_{\alpha,\lambda_n}(0)\right]| \prec_s \frac{1+r_n}{r_n(\lambda_n - r_n \mathtt{L}_\varphi^2)^{3/2}\sqrt{n}} \left(1+\sqrt{\lambda_n - r_n\mathtt{L}_\varphi^2}\right) . \tag{44}$$

which, for $r_n = \lambda_n/(2\mathtt{L}_\varphi^2)$ gives,

$$|\partial_\zeta \xi_{\alpha,\lambda_n}(0) - \mathbb{E}\left[\partial_\zeta \xi_{\alpha,\lambda_n}(0)\right]| \prec_s \frac{1+\lambda_n^2}{\lambda_n^{5/2}\sqrt{n}} .$$

where we stashed the dependancy in $\mathtt{L}_\varphi$ inside the $\prec_s$ notation. We conclude the proof by merging (36), (43) and (44), and using the additivity of $\prec_s$. □

E New deterministic equivalents under correlated data

In this part of the appendix, we state and prove new deterministic equivalents results that generalize the ones of Proposition 1. In particular, Proposition 5, Proposition 6, and Proposition 7 generalize proposition A.4 of [SDCL24] to the context of non-i.i.d. data. This results are the first of their kind, with the exception of [MHD25] which deals with a slightly different setting and only focus on the resolvent matrix. Their result is an hedge case of Proposition 5.

First, we introduce two useful identities, namely the resolvent identity, which states for any two invertible matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$, that,

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}\left\{\mathbf{B} - \mathbf{A}\right\}\mathbf{B}^{-1} . \tag{45}$$

And the leave-two-out trick, which is abstracted as the bellow Lemma 11, and will allow us to desantangle the contribution of a pair of sample from the rest of the data.

Lemma 11. *Let $\mathbf{A} \in \mathbb{R}^{p\times p}$, $\mathbf{U} \in \mathbb{R}^{p\times k}$ and $\mathbf{V} \in \mathbb{R}^{k\times p}$, then,*

$$(\mathbf{A}+\mathbf{U}\mathbf{V})^{-1}\mathbf{U} = \mathbf{A}^{-1}\mathbf{U}\left(\mathrm{I}_k + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}$$

Proof. We have thanks to the Woodburry matrix inversion lemma [HS81] that,

$$\begin{aligned}(\mathbf{A}+\mathbf{U}\mathbf{V})^{-1}\mathbf{U} &= \left(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathrm{I}_k + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}\right)\mathbf{U} \\ &= \mathbf{A}^{-1}\mathbf{U}\left\{\mathrm{I}_k - \left(\mathrm{I}_k + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right\} \\ &= \mathbf{A}^{-1}\mathbf{U}\left(\mathrm{I}_k + \mathbf{V}\mathbf{A}^{-1}\mathbf{U}\right)^{-1} .\end{aligned}$$

□

We consider a variant of the usual deterministic equivalent result of sample covariance resolvents. More precisely, we will consider a sample covariance matrix build from non-i.i.d. samples, in the particular setting where we have i.i.d. pairs $(\Phi_i^{(1)}, \Phi_i^{(2)})$, for $i \in \{1, \dots, n\}$. We define

$$\Phi^{(1)} = (\Phi_i^{(1)})_{i=1}^n \in \mathbb{R}^{p \times k} \ , \quad \text{and} \quad \Phi^{(2)} = (\Phi_i^{(2)})_{i=1}^n \in \mathbb{R}^{p \times n} \ ,$$

and consider the sample covariance matrices $\mathrm{C}^{(1)}$ and $\mathrm{C}^{(2)}$ of $\Phi^{(1)}$ and $\Phi^{(2)}$ respectively defined as,

$$\mathrm{C}^{(1)} = \frac{1}{n} \Phi^{(1)} \Phi^{(1)\top} \in \mathbb{R}^{p \times p} \ , \quad \text{and} \quad \mathrm{C}^{(2)} = \frac{1}{n} \Phi^{(2)} \Phi^{(2)\top} \in \mathbb{R}^{p \times p} \ .$$

as well as the resolvent matrix of $\mathrm{C}^{(1)} + \mathrm{C}^{(2)}$, for any complex matrix $\mathbf{D} \in \mathbb{C}^{p \times p}$ with $\lambda_{\min}^{\Re}(\mathbf{D}) > 0$,

$$\mathrm{R}(\mathbf{D}) = \left(\mathrm{C}^{(1)} + \mathrm{C}^{(2)} + \mathbf{D} \right)^{-1} \ .$$

Previous works such as [Cho22] do not apply in our setting, as $\Phi^{(1)}$ and $\Phi^{(2)}$ are not independent. Perhaps surprisingly, this lack of independence fundamentally alters the structure of the deterministic equivalent of $\mathrm{R}(\mathbf{D})$, rendering the expressions derived in [Cho22] inapplicable. For the sake of brevity, and throughout this appendix, we focus solely on approximating expectations $\mathbb{E}[Q]$ for the quantities Q under consideration, and we do not explicitly address their concentration around the mean. Nevertheless, under a joint concentration assumption on the pairs $(\Phi_i^{(1)}, \Phi_i^{(2)})$ for $i \in \{1, \dots, n\}$—for instance, assuming joint Lipschitz concentration—these quantities naturally concentrate around their expectations. In that case, the proof would follow similar lines to Proposition 4. Finally, we stress that particular effort was devoted to deriving *non-asymptotic bounds* that hold for a finite number of samples n and a finite feature dimension p.

We first prove:

Proposition 5. *Let $\Psi \in \mathbb{R}^{2p \times n}$ be a random matrix with centered, i.i.d., and κ_n-Lipschitz concentrated columns. Let $\mathbf{D} \in \mathbb{C}^{p \times p}$ be a deterministic matrix with $\lambda_{\min}^{\Re}(\mathbf{D}) > 0$. Denote $\Phi^{(1)} = \Psi_{1:p,:} \in \mathbb{R}^{p \times n}$ and $\Phi^{(2)} = \Psi_{p+1:2p,:} \in \mathbb{R}^{p \times n}$, as well as,*

$$\mathrm{R}(\mathbf{D}) = \left(n^{-1} \Phi^{(1)\top} \Phi^{(1)} + n^{-1} \Phi^{(2)\top} \Phi^{(2)} + \mathbf{D} \right)^{-1} \ .$$

Denote $\Phi_1 = (\Phi_1^{(1)}, \Phi_1^{(2)}) \in \mathbb{R}^{p \times 2}$ and write

$$\overline{\mathrm{R}}(\mathbf{D}) = \left(\mathbb{E}\left[\Phi_1 \mathrm{B}(\mathbf{D}) \Phi_1^\top \right] + \mathbf{D} \right)^{-1} \ , \quad \textit{where} \quad \mathrm{B}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1} \mathbb{E}\left[\Phi_1^\top \mathbb{E}\left[\mathrm{R}(\mathbf{D}) \right] \Phi_1 \right] \right)^{-1} \ .$$

Then, it holds for a large enough universal constant c*,*

$$\left\| \mathbb{E}\left[\mathrm{R}(\mathbf{D}) \right] - \overline{\mathrm{R}}(\mathbf{D}) \right\|_{\mathrm{F}} \leqslant \frac{\mathrm{c} \sqrt{p} (1 + \kappa_n^8)}{n \lambda_{\min}^{\Re}(\mathbf{D})^{7/2}} \left\{ \sqrt{\lambda_{\min}^{\Re}(\mathbf{D})} \left(1 + \frac{\sqrt{p}}{n} \right) + \frac{\sqrt{p}}{n} + \frac{\sqrt{p}}{\sqrt{\lambda_{\min}^{\Re}(\mathbf{D}) n}} \right\} \ .$$

Proof. Introduce the following notations,

$$\mathrm{C}^{(1)} = \frac{1}{n} \Phi^{(1)} \Phi^{(1)\top} \in \mathbb{R}^{p \times p} \ , \quad \text{and} \quad \mathrm{C}^{(2)} = \frac{1}{n} \Phi^{(2)} \Phi^{(2)\top} \in \mathbb{R}^{p \times p} \ ,$$

$$\mathrm{C}_-^{(1)} = \mathrm{C}^{(1)} - \frac{1}{n} \Phi_1^{(1)} \Phi_1^{(1)\top} \in \mathbb{R}^{p \times p} \ , \quad \text{and} \quad \mathrm{C}_-^{(2)} = \mathrm{C}^{(2)} - \frac{1}{n} \Phi_1^{(2)} \Phi_1^{(2)\top} \in \mathbb{R}^{p \times p} \ ,$$

$$\Sigma_1 = \mathbb{E}\left[\Phi_1^{(1)} \Phi_1^{(1)\top} \right] \ , \quad \Sigma_2 = \mathbb{E}\left[\Phi_1^{(2)} \Phi_1^{(2)\top} \right] \ , \quad \Sigma_{1,2} = \mathbb{E}\left[\Phi_1^{(1)} \Phi_1^{(2)\top} \right] \ ,$$

and,

$$\mathrm{R}_-(\mathbf{D}) = (\mathrm{C}_-^{(1)} + \mathrm{C}_-^{(2)} + \mathbf{D})^{-1} \ .$$

We first notice that thanks to the above notations, we have,

$$\mathrm{R}(\mathbf{D}) = \left(\mathrm{C}_-^{(1)} + \mathrm{C}_-^{(2)} + \frac{1}{n} \Phi_1 \Phi_1^\top + \mathbf{D} \right)^{-1} \ ,$$

thus, using the matrix inversion Lemma 11, we get,

$$\mathrm{R}(\mathbf{D}) = \mathrm{R}_-(\mathbf{D}) - \frac{1}{n}\mathrm{R}_-(\mathbf{D})\Phi_1 \left(\mathrm{I}_2 + \frac{1}{n}\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1\right)^{-1} \Phi_1^\top \mathrm{R}_-(\mathbf{D}) \,, \tag{46}$$

We will denote $\hat{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1\right)^{-1}$, the dilation matrix that naturally appears in the above application of Lemma 11. Multiplying by Φ_1 in (46), we get,

$$\mathrm{R}(\mathbf{D})\Phi_1 = \mathrm{R}_-(\mathbf{D})\Phi_1\hat{\mathrm{B}}(\mathbf{D}) \,. \tag{47}$$

Leveraging (47), as well as the resolvent identity (45), it results that,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D}) &= \mathbb{E}\left[\mathrm{R}(\mathbf{D})\left\{\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right] - \mathrm{C}^{(1)} - \mathrm{C}^{(2)}\right\}\overline{\mathrm{R}}(\mathbf{D})\right] \\
&= \mathbb{E}\left[\mathrm{R}(\mathbf{D})\left\{\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right] - \Phi_1\Phi_1^\top\right\}\overline{\mathrm{R}}(\mathbf{D})\right] \\
&= \mathbb{E}\left[\mathrm{R}(\mathbf{D}) - \mathrm{R}_-(\mathbf{D})\right]\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D}) \\
&\quad + \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\left\{\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right] - \Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right\}\right]\overline{\mathrm{R}}(\mathbf{D}) \\
&\quad + \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D}) \,,
\end{aligned}$$

which can be further simplified using

$$\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\left\{\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right] - \Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right\}\right] = 0 \,,$$

as well as using (46) again, this proves,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D}) &= \frac{1}{n}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\hat{\mathrm{B}}(\mathbf{D})\Phi_1^\top\mathrm{R}_-(\mathbf{D})\right]\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D}) \\
&\quad + \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D}) \,.
\end{aligned} \tag{48}$$

The remainging of the proof focuses on bounding the Frobenius norms of the two terms in the right-hand side of (48). To this end, we first remark that,

$$\hat{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1\right)^{-1} \preceq \mathrm{I}_2 \,, \quad \text{and} \quad \mathrm{B}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_1^\top\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\Phi_1\right]\right)^{-1} \preceq \mathrm{I}_2 \,.$$

This implies in particular that,

$$\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right] \preceq \mathbb{E}\left[\Phi_1\Phi_1^\top\right] = \Sigma_1 + \Sigma_2 + \Sigma_{1,2} + \Sigma_{1,2}^\top \,.$$

and,

$$\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\hat{\mathrm{B}}(\mathbf{D})\Phi_1^\top\mathrm{R}_-(\mathbf{D})\right] \preceq \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\Phi_1^\top\mathrm{R}_-(\mathbf{D})\right] = \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\left\{\Sigma_1 + \Sigma_2 + \Sigma_{1,2} + \Sigma_{1,2}^\top\right\}\mathrm{R}_-(\mathbf{D})\right] \,,$$

where we have used the tower property and $\preceq$ denotes the Lowner ordering on positive semi-definite matrices. Hence, leveraging Lemma 1, as well as the non-decreasing property of $\|\cdot\|_{\mathrm{op}}$ for $\preceq$, it results that,

$$\|\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right]\|_{\mathrm{op}} \leqslant \|\Sigma_1 + \Sigma_2 + \Sigma_{1,2} + \Sigma_{1,2}^\top\|_{\mathrm{op}} \leqslant \kappa_n^2 \,,$$

and,

$$\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\hat{\mathrm{B}}(\mathbf{D})\Phi_1^\top\mathrm{R}_-(\mathbf{D})\right]\right\|_{\mathrm{op}} \leqslant \frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2} \,.$$

Using these bounds, we can control the norm of the first term in (48), indeed,

$$\frac{1}{n}\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\hat{\mathrm{B}}(\mathbf{D})\Phi_1^\top\mathrm{R}_-(\mathbf{D})\right]\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} \leqslant \frac{\sqrt{p}\kappa_n^4}{n\lambda_{\min}^{\Re}(\mathbf{D})^3} \,. \tag{49}$$

We now focus on the second term in (48), to this end, write,

$$\begin{aligned}
\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} &= \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1} \mathbb{E}\left[\mathrm{tr}\left(\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^\top\overline{\mathrm{R}}(\mathbf{D})\right)\right] \\
&= \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1} \mathbb{E}\left[\mathrm{tr}\left(\Phi_1^\top\overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\right)\right]
\end{aligned} \tag{50}$$

and introduce,

$$\tilde{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + \frac{1}{n}\mathbb{E}\left[\Phi_1 \mathrm{R}_-(\mathbf{D})\Phi_1^\top\right]\right)^{-1} .$$

Then, for any $\mathbf{A}$ such that $\|\mathbf{A}\|_{\mathrm{F}} = 1$, standard calculations yield,

$$\begin{aligned}\mathbb{E}\left[\mathrm{tr}\left(\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1 \left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\right)\right] &= \mathrm{tr}\left(\mathbb{E}\left[\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\right]\left\{\mathrm{B}(\mathbf{D}) - \tilde{\mathrm{B}}(\mathbf{D})\right\}\right) \\ &\quad + \mathbb{E}\left[\mathrm{tr}\left(\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1 \left\{\tilde{\mathrm{B}}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\right)\right] \end{aligned} \tag{51}$$

we control the two terms in (51) , first using the Cauchy-Schwarz inequality,

$$\mathrm{tr}\left(\mathbb{E}\left[\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\right]\left\{\mathrm{B}(\mathbf{D}) - \tilde{\mathrm{B}}(\mathbf{D})\right\}\right) \leqslant \|\mathbb{E}\left[\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\right]\|_{\mathrm{F}} \|\mathrm{B}(\mathbf{D}) - \tilde{\mathrm{B}}(\mathbf{D})\|_{\mathrm{F}} \tag{52}$$

furthermore, we have,

$$\begin{aligned}\|\mathbb{E}\left[\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\right]\|_{\mathrm{F}} &= \left\| \begin{pmatrix} \mathrm{tr}\left(\Sigma_1 \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]\right) & \mathrm{tr}\left(\Sigma_{1,2} \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]\right) \\ \mathrm{tr}\left(\Sigma_{1,2} \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]\right) & \mathrm{tr}\left(\Sigma_2 \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]\right) \end{pmatrix} \right\|_{\mathrm{F}} \\ &\leqslant \frac{2\sqrt{p}\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2} . \end{aligned} \tag{53}$$

and, using the 1-Lipschitz property of $\mathbf{M} \mapsto (\mathrm{I}_2 + \mathbf{M})^{-1}$ on the cone of positive semi-definite matrices, as well as Equation (46), we have that,

$$\begin{aligned}\|\mathrm{B}(\mathbf{D}) - \tilde{\mathrm{B}}(\mathbf{D})\|_{\mathrm{F}} &\leqslant \frac{1}{n}\left\|\mathbb{E}\left[\Phi_1^\top \mathbb{E}\left[\mathrm{R}(\mathbf{D}) - \mathrm{R}_-(\mathbf{D})\right]\Phi_1\right]\right\|_{\mathrm{F}} \leqslant \frac{1}{n^2}\left\|\mathbb{E}\left[\Phi_1^\top \mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1 \hat{\mathrm{B}}(\mathbf{D})\Phi_1^\top \mathrm{R}_-(\mathbf{D})\right]\Phi_1\right]\right\|_{\mathrm{F}} \\ &\leqslant \frac{1}{n^2}\left\|\mathbb{E}\left[\Phi_1^\top \mathrm{R}_-(\mathbf{D})\mathbb{E}\left[\Phi_1\Phi_1^\top\right]\mathrm{R}_-(\mathbf{D})\Phi_1\right]\right\|_{\mathrm{F}} \leqslant \frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n^2} . \end{aligned} \tag{54}$$

Plugging (53) and (54) in (52), we get,

$$\mathrm{tr}\left(\mathbb{E}\left[\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\right]\left\{\mathrm{B}(\mathbf{D}) - \tilde{\mathrm{B}}(\mathbf{D})\right\}\right) \leqslant \frac{2\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n^2} . \tag{55}$$

Moving on to the second term in the right hand side of (51), we introduce

$$\chi(\mathbf{D}) = \Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1 ,$$

so that, using (45) yields

$$\begin{aligned}&\mathbb{E}\left[\mathrm{tr}\left(\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1 \left\{\tilde{\mathrm{B}}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\right)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\mathrm{tr}\left(\hat{\mathrm{B}}(\mathbf{D})\chi(\mathbf{D})\tilde{\mathrm{B}}(\mathbf{D})\left\{\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1 - \mathbb{E}\left[\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1\right]\right\}\right)\right] \\ &\leqslant \frac{1}{n}\sqrt{\mathrm{Var}(\hat{\mathrm{B}}(\mathbf{D})\chi(\mathbf{D}))\mathrm{Var}(\Phi_1 \mathrm{R}_-(\mathbf{D})\Phi_1^\top)} . \end{aligned} \tag{56}$$

Note that the last line results from the Cauchy-Schwarz inequality, as well as $\tilde{\mathrm{B}}(\mathbf{D})$ being a deterministic matrix with $\|\tilde{\mathrm{B}}(\mathbf{D})\|_{\mathrm{op}} \leqslant 1$, the Var operator being defined for random matrices in (24). We bound the variances of each entry of $\Phi_1 \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1^\top$ thanks to Corollary 1 which gives,

$$\mathrm{Var}(\Phi_1 \mathrm{R}_-(\mathbf{D})\Phi_1^\top) \leqslant \mathrm{c}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^3}\right) \tag{57}$$

and, applying Lemma 3 to the variance of $\hat{\mathrm{B}}(\mathbf{D})\chi(\mathbf{D})$, we get,

$$\mathrm{Var}(\hat{\mathrm{B}}(\mathbf{D})\chi(\mathbf{D})) \leqslant 2\mathrm{Var}(\chi(\mathbf{D})) + 2\|\mathbb{E}\left[\chi(\mathbf{D})\right]\|_{\mathrm{op}}^2 \mathrm{Var}(\hat{\mathrm{B}}(\mathbf{D})) . \tag{58}$$

additionally the 1-Lipschitz property of $\mathbf{M} \mapsto (\mathrm{I}_2 + \mathbf{M})^{-1}$ on the space of positive matrices, implies

$$\mathrm{Var}(\hat{\mathrm{B}}(\mathbf{D})) \leqslant \frac{1}{n^2}\mathrm{Var}(\Phi_1^\top \mathrm{R}_-(\mathbf{D})\Phi_1) \leqslant \frac{\mathrm{c}}{n^2}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^3}\right) . \tag{59}$$

Finally, using Lemma 1 and Corollary 1, we have that,

$$\|\mathbb{E}\left[\chi(\mathbf{D})\right]\|_{\mathrm{op}} \leqslant \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\left(\|\Sigma_1\|_{\mathrm{op}} + \|\Sigma_2\|_{\mathrm{op}} + \|\Sigma_{1,2}\|_{\mathrm{op}} + \|\Sigma_{1,2}^\top\|_{\mathrm{op}}\right) \leqslant \frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2} , \tag{60}$$

and,

$$\mathrm{Var}(\chi(\mathbf{D})) \leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^4} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^5}\right) . \tag{61}$$

To conclude, we have by merging (57), (58), (59), (60) and (61), we get,

$$\mathrm{Var}(\hat{\mathrm{B}}(\mathbf{D})\chi(\mathbf{D})) \leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^4} + \frac{\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^5 n} + \frac{\kappa_n^8 p}{\lambda_{\min}^{\Re}(\mathbf{D})^6 n^2} + \frac{\kappa_n^{10} p}{\lambda_{\min}^{\Re}(\mathbf{D})^7 n^3}\right) . \tag{62}$$

And, plugging (57) and (62) in (56), we get for a large enough constant c,

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{tr}\left(\Phi_1^\top \overline{\mathrm{R}}(\mathbf{D})\mathbf{A}\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\tilde{\mathrm{B}}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\right)\right] \qquad (63)\\
&\leqslant \frac{\mathrm{c}}{n}\left(\frac{\kappa_n^2\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})} + \frac{\sqrt{p}\kappa_n^3}{\sqrt{n}\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}}\right)\left(\frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{\kappa_n^3\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2}\sqrt{n}} + \frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n} + \frac{\kappa_n^5\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2}n^{3/2}}\right)\\
&= \frac{\mathrm{c}\sqrt{p}}{n}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^3} + \frac{2\kappa_n^5\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}\sqrt{n}} + \frac{\kappa_n^6\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n} + \frac{2\kappa_n^7\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}n^{3/2}} + \frac{\kappa_n^6\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n} + \frac{\kappa_n^8\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n^2}\right) .
\end{aligned}$$

By plugging (55), (63) and (51) in (50), we thus get,

$$\begin{aligned}
&\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1\left\{\mathrm{B}(\mathbf{D}) - \hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^\top\right]\overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} \qquad (64)\\
&\leqslant \frac{\mathrm{c}\sqrt{p}}{n}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^3} + \frac{\kappa_n^5\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}\sqrt{n}} + \frac{\kappa_n^6\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n} + \frac{\kappa_n^7\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}n^{3/2}}\right.\\
&\qquad\left. + \frac{\kappa_n^6\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n} + \frac{\kappa_n^8\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n^2}\right) .
\end{aligned}$$

The claim follows from (48), (49) and(64), as well as merely reorganizing each term. □

We derive similar results for some variants of $\mathrm{R}(\mathbf{D})$, which directly relate to Proposition 1. Indeed, the following Proposition 6can be readily applied to approximate $\theta_\star^\top \Sigma_\star \mathbb{E}\left[\hat{\theta}_{\alpha,\lambda_n}\right]$, we show:

Proposition 6. *Let $\Psi \in \mathbb{R}^{(2p+2)\times n}$ be a random matrix with centered, i.i.d., and κ_n-Lipschitz concentrated columns. Let $\mathbf{D} \in \mathbb{C}^{p\times p}$ be a deterministic matrix with $\lambda_{\min}^{\Re}(\mathbf{D}) > 0$. Denote $\Phi^{(1)} = \Psi_{1:p,:} \in \mathbb{R}^{p\times n}$, $\Phi^{(2)} = \Psi_{p+1:2p,:} \in \mathbb{R}^{p\times n}$, and $\Phi^{(3)} = \Psi_{2p+1,:} \in \mathbb{R}^{1\times n}$, and $\Phi^{(4)} = \Psi_{2p+2,:} \in \mathbb{R}^{1\times n}$ as well as for all $(i,j) \in \{(1,2),(3,4)\}$, $\Phi^{(i,j)} = (\Phi^{(i)}, \Phi^{(j)})$ and $\Phi_1^{(i,j)} = (\Phi_1^{(i)}, \Phi_1^{(j)})$, we consider*

$$\mathrm{R}(\mathbf{D}) = \left(n^{-1}\Phi^{(1,2)}\Phi^{(1,2)\top} + \mathbf{D}\right)^{-1} , \quad \mathrm{Q}_{\mathrm{Aug}}(\mathbf{D}) = \mathrm{R}(\mathbf{D})n^{-1}\Phi^{(1,2)}\Phi^{(3,4)\top} .$$

Write,

$$\overline{\mathrm{R}}(\mathbf{D}) = \left(\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)\top}\right] + \mathbf{D}\right)^{-1} , \quad \textit{where} \quad \mathrm{B}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\right)^{-1} ,$$

define,

$$\overline{\mathrm{Q}}_{\mathrm{Aug}}(\mathbf{D}) = \overline{\mathrm{R}}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right] .$$

Then, it holds for a large enough universal constant c*,*

$$\begin{aligned}\left\|\mathbb{E}\left[\mathrm{Q}_{\mathrm{Aug}}(\mathbf{D})\right]-\overline{\mathrm{Q}}_{\mathrm{Aug}}(\mathbf{D})\right\|_2 \leqslant \kappa_n^2\left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]-\overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} \\ + \frac{\mathrm{c}\sqrt{p}(1+\kappa_n^{13})}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n}\left\{\lambda_{\min}^{\Re}(\mathbf{D})+\lambda_{\min}^{\Re}(\mathbf{D})^2+\frac{p}{n}+\frac{p}{\lambda_{\min}^{\Re}(\mathbf{D})n^2}+\frac{p}{\lambda_{\min}^{\Re}(\mathbf{D})^2n^3}+\frac{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}}{\sqrt{n}}\right. \\ \left.+\frac{p}{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}n^{3/2}}+\frac{p}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}n^{5/2}}+\frac{p}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2}n^{7/2}}\right\},\end{aligned}$$

Proof. We first introduce the following matrices,

$$\mathrm{R}_-(\mathbf{D})=\left(n^{-1}\Phi^{(1,2)}\Phi^{(1,2)\top}-n^{-1}\Phi_1^{(1,2)}\Phi_1^{(1,2)\top}+\mathbf{D}\right)^{-1}$$

for all $i\in\{1,2\}$,

$$\mathrm{C}^{(i)}=\frac{1}{n}\Phi^{(i)\top}\Phi^{(i)}\ ,\quad \mathrm{C}_-^{(i)}=\mathrm{C}^{(i)}-\frac{1}{n}\Phi_1^{(i)}\Phi_1^{(i)\top}\ ,\quad \text{and}\quad \Sigma_i=\mathbb{E}\left[\Phi_1^{(i)}\Phi_1^{(i)\top}\right]$$

and for all $(i,j)\in\{(1,2),(3,4)\}$,

$$\mathrm{C}^{(i,j)}=\frac{1}{n}\Phi^{(i)\top}\Phi^{(j)}\ ,\quad \mathrm{C}_-^{(i,j)}=\mathrm{C}^{(i,j)}-\frac{1}{n}\Phi_1^{(i)}\Phi_1^{(j)\top}\ ,\quad \text{and}\quad \Sigma_{i,j}=\mathbb{E}\left[\Phi_1^{(i)}\Phi_1^{(j)\top}\right]$$

as well as the dilation matrix,

$$\hat{\mathrm{B}}(\mathbf{D})=\left(\mathrm{I}_2+n^{-1}\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right)^{-1}\ .$$

Thanks to calculations similar to the ones presented in the proof of Proposition 5, we write,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Q}_{\mathrm{Aug}}(\mathbf{D})\right]&=\mathbb{E}\left[\mathrm{R}(\mathbf{D})\Phi_1^{(1,2)}\Phi_1^{(3,4)\top}\right]=\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\\&=\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D})-\mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]+\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\ ,\end{aligned}$$

hence,

$$\begin{aligned}\mathbb{E}\left[\mathrm{Q}_{\mathrm{Aug}}(\mathbf{D})\right]-\overline{\mathrm{Q}}_{\mathrm{Aug}}(\mathbf{D})&=\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D})-\mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\\&+\left\{\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]-\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\right\}\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\\&+\left\{\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]-\overline{\mathrm{R}}(\mathbf{D})\right\}\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\ .\end{aligned}$$

Recalling that $\mathrm{B}(\mathbf{D})\preceq\mathrm{I}_2$, and using Lemma 1,we have,

$$\left\|\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\right\|_{\mathrm{op}}\leqslant\left\|\mathbb{E}\left[\Phi_1^{(1,2)}\Phi_1^{(3,4)\top}\right]\right\|_{\mathrm{op}}\leqslant\kappa_n^2\ ,$$

Additionally, we have from the resolvent identity, and Lemma 1 again,

$$\begin{aligned}\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]-\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\right\|_{\mathrm{F}}&=\frac{1}{n}\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\right]\right\|_{\mathrm{F}}\\&\leqslant\frac{\kappa_n^2\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n}\ ,\end{aligned}$$

thus,

$$\left\|\left\{\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\right]-\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\right\}\mathbb{E}\left[\Phi_1\mathrm{B}(\mathbf{D})\Phi_1^{(3)}\right]\right\|_{\mathrm{F}}\leqslant\frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n}\ ,$$

and,

$$\begin{aligned}\left\|\mathbb{E}\left[\mathrm{Q}_{\mathrm{Aug}}(\mathbf{D})\right]-\overline{\mathrm{Q}}_{\mathrm{Aug}}(\mathbf{D})\right\|_2\leqslant\ &\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D})-\mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_2\\&+\frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n}+\kappa_n^2\left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]-\overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}}\end{aligned}\tag{65}$$

It remains to control the first term in the right-hand side of (65), to this end, we introduce the notation $\widetilde{\mathrm{B}}(\mathbf{D})$ for,

$$\widetilde{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right]\right)^{-1} ,$$

then we have,

$$\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_2 \leqslant \left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_2 + \frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\|\mathrm{B}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\|_{\mathrm{op}} .$$

Proceeding similarly to (54), we have,

$$\|\mathrm{B}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\|_{\mathrm{op}} \leqslant \frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n^2} ,$$

and from the resolvent identity,

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_2 \\ &= \frac{1}{n}\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\left\{\Phi_1^{(1,2)^\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)} - \mathbb{E}\left[\Phi_1^{(1,2)^\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right]\right\}\widetilde{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\right\|_2 .\end{aligned}$$

Thus, using $\|\mathbf{x}\|_2 = \sup_{\|\mathbf{a}\|_2=1}\mathbf{a}^\top\mathbf{x}$ and the Cauchy-Schwarz inequality, we have,

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_2 \\ &\leqslant \sup_{\|\mathbf{a}\|_2=1}\frac{1}{n}\sqrt{\mathrm{Var}\left(\widetilde{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\right)\mathrm{Var}\left(\Phi_1^{(1,2)^\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right)} \\ &\leqslant \frac{\mathrm{c}\sqrt{p}}{n}\left(\frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})} + \frac{\kappa_n^3}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}\sqrt{n}}\right)\sup_{\|\mathbf{a}\|_2=1}\sqrt{\mathrm{Var}\left(\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\right)} . \qquad (66)\end{aligned}$$

The last line follows from the same lines as (57), and the fact that $\widetilde{\mathrm{B}}(\mathbf{D})$ is deterministic and $\widetilde{\mathrm{B}}(\mathbf{D}) \preceq \mathrm{I}_2$. Now, fix $\mathbf{a}$ such that $\|\mathbf{a}\|_2 = 1$, we write for simplicity,

$$\chi(\mathbf{D}) = \Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)} ,$$

then we have from Lemma 3,

$$\mathrm{Var}\left(\chi(\mathbf{D})\hat{\mathrm{B}}(\mathbf{D})\right) \leqslant \mathrm{Var}\left(\chi(\mathbf{D})\right) + \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\right) ,$$

then, similarly to (59), we have,

$$\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\right) \leqslant \frac{\mathrm{c}}{n^2}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^3}\right) ,$$

and using Corollary 1 on all the entries of $\chi(\mathbf{D})$, then over the entries, it results,

$$\mathrm{Var}\left(\chi(\mathbf{D})\right) \leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^3}\right) ,$$

and finally,

$$\mathrm{Var}\left(\chi(\mathbf{D})\hat{\mathrm{B}}(\mathbf{D})\right) \leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{p\kappa_n^6}{n\lambda_{\min}^{\Re}(\mathbf{D})^3}\right) + \frac{\mathrm{c}p}{n^2}\left(\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^4} + \frac{\kappa_n^{10}}{n\lambda_{\min}^{\Re}(\mathbf{D})^5}\right) . \qquad (67)$$

Plugging (67) in (66), we get,

$$\begin{aligned}&\left\|\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \widetilde{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right\|_{\mathrm{F}} \qquad (68)\\ &\leqslant \frac{\mathrm{c}\sqrt{p}}{n}\Bigg(\frac{\kappa_n^6}{\lambda_{\min}^{\Re}(\mathbf{D})^3} + \frac{\kappa_n^8 p}{\lambda_{\min}^{\Re}(\mathbf{D})^4 n} + \frac{\kappa_n^{10} p}{\lambda_{\min}^{\Re}(\mathbf{D})^5 n^2} + \frac{\kappa_n^{12} p}{\lambda_{\min}^{\Re}(\mathbf{D})^6 n^3} + \frac{\kappa_n^7}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}\sqrt{n}} \\ &\qquad + \frac{\kappa_n^9 p}{\lambda_{\min}^{\Re}(\mathbf{D})^{9/2} n^{3/2}} + \frac{\kappa_n^{11} p}{\lambda_{\min}^{\Re}(\mathbf{D})^{11/2} n^{5/2}} + \frac{\kappa_n^{13} p}{\lambda_{\min}^{\Re}(\mathbf{D})^{13/2} n^{7/2}}\Bigg) .\end{aligned}$$

The result follows from plugging (68) in (65), as well as merely reorganizing each term. □

Using Proposition 6 with $\Psi = (\Psi_i)_{i=1}^n$, s.t. for all $i \leqslant n$, $\Psi_i = (\sqrt{1-\alpha}\varphi(\mathtt{X}_i), \sqrt{\alpha}\mu_x(\mathtt{Z}_i), \sqrt{1-\alpha}\mathtt{Y}_i, \sqrt{\alpha}\mu_y(\mathtt{Z}_i))$ and $\mathbf{D} = \alpha\mathbb{E}[\Lambda(\mathtt{Z})] + \lambda_n \mathrm{I}_p$, one can approximate $\mathbb{E}\left[\hat{\theta}_{\alpha,\lambda}\right]$. This will be formalized later in Section F. In the meantime, we prove a comparable result that allows to approximate $\mathbb{E}[\xi_{\alpha,\lambda}(\zeta)]$ defined in (35), for any $\zeta \in \mathsf{U}_\lambda$:

Proposition 7. *Let $\Psi \in \mathbb{R}^{(2p+2)\times n}$ be a random matrix with centered i.i.d., and κ_n Lipschitz concentrated columns. Let $\mathbf{D} \in \mathbb{C}^{p\times p}$ be a deterministic matrix with $\lambda^{\Re}_{\min}(\mathbf{D}) > 0$. Denote $\Phi^{(1)} = \Phi_{1:p,:} \in \mathbb{R}^{p\times n}$, $\Phi^{(2)} = \Phi_{p+1:2p,:} \in \mathbb{R}^{p\times n}$, $\Phi^{(3)} = \Psi_{2p+1,:} \in \mathbb{R}^{1\times n}$ and $\Phi^{(4)} = \Psi_{2p+2,:} \in \mathbb{R}^{1\times n}$, as well as for all $(i,j) \in \{(1,2),(3,4)\}$, $\Phi^{(i,j)} = (\Phi^{(i)}, \Phi^{(j)})$ and for $\Phi_1^{(i,j)} = (\Phi_1^{(i)}, \Phi_1^{(j)})$, we consider*

$$\mathrm{R}(\mathbf{D}) = \left(n^{-1}\Phi^{(1,2)}\Phi^{(1,2)\top} + \mathbf{D}\right)^{-1}, \quad \mathrm{S}_{\mathrm{Aug}}(\mathbf{D}) = n^{-1}\Phi^{(3,4)}\Phi^{(1,2)\top}\mathrm{R}(\mathbf{D})n^{-1}\Phi^{(1,2)}\Phi^{(3,4)\top}.$$

Define,

$$\overline{\mathrm{R}}(\mathbf{D}) = \left(\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)\top}\right] + \mathbf{D}\right)^{-1}, \quad \textit{where} \quad \mathrm{B}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_1^{(1,2)^\top}\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\right)^{-1},$$

and,

$$\overline{\mathrm{S}}_{\mathrm{Aug}}(\mathbf{D}) = \mathbb{E}\left[\Phi_1^{(3,4)}\{\mathrm{I}_2 - \mathrm{B}(\mathbf{D})\}\Phi_1^{(3,4)\top}\right] + \mathbb{E}\left[\Phi_1^{(3,4)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)\top}\right]\overline{\mathrm{R}}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)}\right].$$

Then, it holds for a large enough constant c*,*

$$\begin{aligned}
\left|\mathbb{E}\left[\mathrm{S}_{\mathrm{Aug}}(\mathbf{D})\right] - \overline{\mathrm{S}}_{\mathrm{Aug}}(\mathbf{D})\right| \leqslant \mathrm{c}\kappa_n^2 \left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} + \frac{\mathrm{c}(1+\kappa_n^7)\sqrt{p}}{\lambda^{\Re}_{\min}(\mathbf{D})^{5/2}n}\left(\lambda^{\Re}_{\min}(\mathbf{D})^{1/2} + \lambda^{\Re}_{\min}(\mathbf{D}) + \frac{\sqrt{p}}{\sqrt{n}}\right) \\
+ \frac{\mathrm{c}\kappa_n^8 p^{3/2}}{\lambda^{\Re}_{\min}(\mathbf{D})^3 n^2} + \frac{\mathrm{c}(1+\kappa_n^8)\sqrt{p}}{\lambda^{\Re}_{\min}(\mathbf{D})^3 n^{3/2}}\left(\lambda^{\Re}_{\min}(\mathbf{D})^{1/2} + \frac{\sqrt{p}}{\sqrt{n}}\right) + \frac{\mathrm{c}\kappa_n^4}{\lambda^{\Re}_{\min}(\mathbf{D})n} \\
+ \frac{\mathrm{c}\kappa_n^6 p}{\lambda^{\Re}_{\min}(\mathbf{D})^2 n^2} + \frac{\mathrm{c}(1+\kappa_n^9)p}{\lambda^{\Re}_{\min}(\mathbf{D})^{7/2}n^{5/2}}\left(1 + \lambda^{\Re}_{\min}(\mathbf{D})^{1/2} + \lambda^{\Re}_{\min}(\mathbf{D})^{3/2}\right) \\
+ \frac{\mathrm{c}(1+\kappa_n^{10})p}{\lambda^{\Re}_{\min}(\mathbf{D})^4 n^3}\left(1 + \lambda^{\Re}_{\min}(\mathbf{D})\right) + \frac{\mathrm{c}(1+\kappa_n^9)\sqrt{p}}{\lambda^{\Re}_{\min}(\mathbf{D})^{7/2}n^{5/2}}\left(1 + \lambda^{\Re}_{\min}(\mathbf{D})^{1/2}+\right)
\end{aligned}$$

Proof. We first introduce the following matrices,

$$\mathrm{R}_-(\mathbf{D}) = \left(n^{-1}\Phi^{(1,2)}\Phi^{(1,2)\top} - n^{-1}\Phi_1^{(1,2)}\Phi_1^{(1,2)\top} + \mathbf{D}\right)^{-1},$$

$$\mathrm{R}_{--}(\mathbf{D}) = \left(n^{-1}\Phi^{(1,2)}\Phi^{(1,2)\top} - n^{-1}\Phi_1^{(1,2)}\Phi_1^{(1,2)\top} - n^{-1}\Phi_2^{(1,2)}\Phi_2^{(1,2)\top} + \mathbf{D}\right)^{-1},$$

where for all $(i,j) \in \{1,2,3,4\}^2$, $\Phi_2^{(i,j)} = (\Phi_2^{(i)}, \Phi_2^{(j)})$ and

$$\Sigma_i = \mathbb{E}\left[\Phi_1^{(i)}\Phi_1^{(i)\top}\right], \quad \Sigma_{i,j} = \mathbb{E}\left[\Phi_1^{(i)}\Phi_1^{(j)\top}\right],$$

as well as the matrix empirical dilation factors,

$$\hat{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right)^{-1}.$$

Using the fact that the rows of Ψ have identical distribution, we start by writting,

$$\mathbb{E}\left[\mathrm{S}_{\mathrm{Aug}}(\mathbf{D})\right] = \underbrace{\frac{1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\Phi_1^{(1,2)\top}\mathrm{R}(\mathbf{D})\Phi_1^{(1,2)}\Phi_1^{(3,4)\top}\right]}_{s_1} + \underbrace{\frac{n(n-1)}{n^2}\mathbb{E}\left[\Phi_1^{(3,4)}\Phi_1^{(1,2)\top}\mathrm{R}(\mathbf{D})\Phi_2^{(1,2)}\Phi_2^{(3,4)\top}\right]}_{s_2}.$$

In the following, we deal with the terms s_1 and s_2 separately. First, we focus on s_1 and write, thanks to (47),

$$\begin{aligned} s_1 &= \frac{1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\Phi_1^{(1,2)\top}\mathrm{R}(\mathbf{D})\Phi_1^{(1,2)}\Phi_1^{(3,4)\top}\right] \\ &= \mathbb{E}\left[\Phi_1^{(3,4)}\frac{1}{n}\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right] \\ &= \mathbb{E}\left[\Phi_1^{(3,4)}\left\{\mathrm{I}_2-\hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right] \\ &= \underbrace{\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\mathrm{I}_2-\mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]}_{\overline{s}_1} + \mathbb{E}\left[\Phi_1^{(3,4)}\left\{\mathrm{B}(\mathbf{D})-\hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right] . \end{aligned} \tag{69}$$

We introduce for notational convenience,

$$\widetilde{\mathrm{B}}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right]\right)^{-1} .$$

The previous calculation in (69) combined with the triangle inequality yields,

$$\begin{aligned} |s_1-\overline{s}_1| &\leqslant \left|\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\widetilde{\mathrm{B}}(\mathbf{D})-\mathrm{B}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right| + \left|\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\widetilde{\mathrm{B}}(\mathbf{D})-\hat{\mathrm{B}}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\right]\right| \\ &\leqslant \frac{1}{n^2}\left|\mathbb{E}\left[\Phi_1^{(3,4)}\widetilde{\mathrm{B}}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\mathrm{B}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\right| \\ &\quad + \frac{1}{n}\left|\mathbb{E}\left[\Phi_1^{(3,4)}\widetilde{\mathrm{B}}(\mathbf{D})\left\{\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)} - \mathbb{E}\left[\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right]\right\}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\right| \end{aligned}$$

where the last inequality follows from (45) and (47). Further rewritting the upper bound, and using the Cauchy-Schwarz inequality, we get,

$$\begin{aligned} |s_1-\overline{s}_1| &\leqslant \frac{1}{n^2}\left|\mathrm{tr}\left(\mathrm{B}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right]\widetilde{\mathrm{B}}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\right)\right| \\ &\quad + \frac{1}{n}\left|\mathrm{tr}\left(\mathbb{E}\left[\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\widetilde{\mathrm{B}}(\mathbf{D})\left\{\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)} - \mathbb{E}\left[\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right]\right\}\right]\right)\right| \\ &\leqslant \frac{2\kappa_n^2}{n^2}\left\|\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\right\|_{\mathrm{F}} \\ &\quad + \frac{1}{n}\sqrt{\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\widetilde{\mathrm{B}}(\mathbf{D})\right)\mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right)} , \end{aligned} \tag{70}$$

where we have used Lemma 1, as well as $\widetilde{\mathrm{B}}(\mathbf{D}) \preceq \mathrm{I}_2$ and $\mathrm{B}(\mathbf{D}) \preceq \mathrm{I}_2$. Additionaly,

$$\left\|\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\right\|_{\mathrm{F}} \leqslant \frac{4\sqrt{p}\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2} ,$$

and, $\widetilde{\mathrm{B}}(\mathbf{D})$ being deterministic with $\widetilde{\mathrm{B}}(\mathbf{D}) \preceq \mathrm{I}_2$, we have,

$$\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\widetilde{\mathrm{B}}(\mathbf{D})\right) \leqslant \mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right) .$$

Thus, (70) implies,

$$|s_1-\overline{s}_1| \leqslant \frac{8\sqrt{p}\kappa_n^6}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n^2} + \frac{1}{n}\sqrt{\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right)\mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right)} . \tag{71}$$

By similar lines as in (57) we get,

$$\mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right) \leqslant \mathrm{c}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right) , \tag{72}$$

finally, using Lemma 3 and Corollary 1, we get,

$$\begin{aligned} \mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right) &\leqslant \mathrm{Var}\left(\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right) + \kappa_n^4\mathrm{Var}\left(\hat{\mathrm{B}}(\mathbf{D})\right) \\ &\leqslant \mathrm{Var}\left(\Phi_1^{(3,4)\top}\Phi_1^{(3,4)}\right) + \frac{\kappa_n^4}{n^2}\mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\right) \\ &\leqslant 2\kappa_n^4 + \frac{\mathrm{c}}{n^2}\left(\frac{\kappa_n^8 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{\kappa_n^{10} p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right) . \end{aligned} \tag{73}$$

Plugging (72) and (73) into (71), we get for a large enough constant c,

$$|s_1 - \bar{s}_1| \leqslant \frac{\mathrm{c}\kappa_n^4\sqrt{p}}{\lambda^{\Re}_{\min}(\mathbf{D})^2 n} + \frac{\mathrm{c}\kappa_n^5\sqrt{p}}{\lambda^{\Re}_{\min}(\mathbf{D})^{3/2} n^{3/2}} + \frac{\mathrm{c}p}{n^2}\left(\frac{\kappa_n^6}{\lambda^{\Re}_{\min}(\mathbf{D})^2} + \frac{\kappa_n^7}{\lambda^{\Re}_{\min}(\mathbf{D})^{5/2}\sqrt{n}}\right) + \frac{\mathrm{c}\kappa_n^8 p}{\lambda^{\Re}_{\min}(\mathbf{D})^3 n^3} \,. \tag{74}$$

Moving on to the second term s_2, we denote by $\mathcal{F}_- = \sigma(\Psi_i \,:\, i \in \{2,\dots,n\})$ and $\mathcal{F}_{--} = \sigma(\Psi_i \,:\, i \in \{3,\dots,n\})$ the σ-algebra generated by Ψ but its first two columns, as well as the following dilation matrices where two samples were removed,

$$\hat{\mathrm{B}}_{-i}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\Phi_i^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_i^{(1,2)}\right)^{-1} \,, \quad \widetilde{\mathrm{B}}_{-i}(\mathbf{D}) = \left(\mathrm{I}_2 + n^{-1}\mathbb{E}\left[\Phi_i^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_i^{(1,2)}\right]\right)^{-1}$$

We then have from Lemma 11 and the tower property,

$$\begin{aligned}
s_2 &= \frac{n-1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\Phi_1^{(1,2)\top}\mathrm{R}(\mathbf{D})\Phi_2^{(1,2)}\Phi_2^{(3,4)\top}\right] \\
&= \frac{n-1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right] \\
&= \tilde{s}_2^1 + \frac{n-1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\hat{\mathrm{B}}(\mathbf{D}) - \hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right] \,,
\end{aligned}$$

where we write for ease of notation

$$\tilde{s}_2^1 = \frac{n-1}{n}\mathbb{E}\left[\mathbb{E}\left[\Phi_1^{(3,4)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \,\middle|\, \mathcal{F}_{--}\right]\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \,\middle|\, \mathcal{F}_{--}\right]\right] \,.$$

Then, using (45), we have,

$$\begin{aligned}
&\left|s_2 - \tilde{s}_2^1\right| \\
&\leqslant \frac{1}{n}\left|\mathbb{E}\left[\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\left\{\mathrm{R}_{--}(\mathbf{D}) - \mathrm{R}_-(\mathbf{D})\right\}\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right]\right| \\
&\leqslant \frac{1}{n^2}\left|\mathbb{E}\left[\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\Phi_2^{(1,2)\top}\mathrm{R}_-(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right]\right| \\
&\leqslant \frac{1}{n^2}\left|\mathbb{E}\left[\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right]\right|
\end{aligned}$$

where we used Lemma 11 in the last inequality. Further writting

$$\begin{aligned}
\chi_1(\mathbf{D}) &= \Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top} \,, \\
\chi_2(\mathbf{D}) &= \Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \,,
\end{aligned}$$

we have from the previous as well as the Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|s_2 - \tilde{s}_2^1\right| &\leqslant \frac{1}{n^2}\left|\mathbb{E}\left[\chi_1(\mathbf{D})^\top\mathrm{R}_{--}(\mathbf{D})\chi_2(\mathbf{D})\right]\right| \\
&\leqslant \frac{1}{n^2}\left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2\left\|\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right]\right\|_{\mathrm{op}}\left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2 \\
&\quad + \frac{1}{n^2}\left(\sqrt{\mathrm{Var}\left(\chi_1(\mathbf{D})\right)\mathrm{Var}\left(\mathrm{R}_{--}(\mathbf{D})\chi_2(\mathbf{D})\right)}\right. \\
&\quad \left. + \left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2\sqrt{\mathbb{E}\left[\left\|\mathrm{R}_{--}(\mathbf{D}) - \mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right]\right\|_{\mathrm{op}}^2\right]\mathrm{Var}\left(\chi_2(\mathbf{D})\right)}\right) \,.
\end{aligned}$$

In order to bound this error, we first remark that,

$$\left\|\mathrm{R}_{--}(\mathbf{D})\right\|_{\mathrm{op}} \leqslant 1/\lambda^{\Re}_{\min}(\mathbf{D}) \,, \qquad \mathbb{E}\left[\left\|\mathrm{R}_{--}(\mathbf{D}) - \mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right]\right\|_{\mathrm{op}}^2\right] \leqslant \mathrm{c}/\lambda^{\Re}_{\min}(\mathbf{D})^2 \,.$$

Furthermore, we have from lines simialar as Lemma 3,

$$\begin{aligned}
\mathrm{Var}\left(\mathrm{R}_{--}(\mathbf{D})\chi_2(\mathbf{D})\right) &\leqslant \frac{2}{\lambda^{\Re}_{\min}(\mathbf{D})^2}\mathrm{Var}\left(\chi_2(\mathbf{D})\right) + 2\left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2\mathbb{E}\left[\left\|\mathrm{R}(\mathbf{D}) - \mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\right\|_{\mathrm{op}}^2\right] \\
&\leqslant \frac{2}{\lambda^{\Re}_{\min}(\mathbf{D})^2}\left(\mathrm{Var}\left(\chi_2(\mathbf{D})\right) + \left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2^2\right) \,,
\end{aligned}$$

where we have used

$$\mathbb{E}\left[\|\mathrm{R}(\mathbf{D}) - \mathbb{E}\left[\mathrm{R}(\mathbf{D})\right]\|_{\mathrm{op}}^2\right] \leqslant \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^2} \,.$$

We have shown,

$$\begin{aligned}\left|s_2 - \tilde{s}_2^1\right| \leqslant \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})n^2}\left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2\left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2 + \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})n^2}\sqrt{\mathrm{Var}(\chi_1(\mathbf{D}))\mathrm{Var}(\chi_2(\mathbf{D}))} \\ + \frac{\left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2}{\lambda_{\min}^{\Re}(\mathbf{D})n^2}\sqrt{\mathrm{Var}\left(\chi_1(\mathbf{D})\right)} + \frac{\mathrm{c}\left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2}{\lambda_{\min}^{\Re}(\mathbf{D})n^2}\sqrt{\mathrm{Var}\left(\chi_2(\mathbf{D})\right)} \,. \end{aligned} \tag{75}$$

In addition, we bound the first two moments of $\chi_1(\mathbf{D})$ and $\chi_2(\mathbf{D})$, first by conditionning on $\mathcal{F}_{--}$ and using the tower property, Lemma 1, we write

$$\begin{aligned}\left\|\mathbb{E}\left[\chi_2(\mathbf{D})\right]\right\|_2 &\leqslant \mathbb{E}\left[\left\|\mathbb{E}\left[\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \;\middle|\; \mathcal{F}_{--}\right]\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \;\middle|\; \mathcal{F}_{--}\right]\right\|_2\right] \\ &\leqslant \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})} \,, \end{aligned} \tag{76}$$

similarly for the expectation of $\chi_1(\mathbf{D})$, write,

$$\begin{aligned}\left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2 &\leqslant \left\|\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\right]\mathbb{E}\left[\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right]\right\|_2 \\ &\quad + \sup_{\|\mathbf{a}\|_2=1}\sqrt{\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\right)\mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top}\right)} \\ &\leqslant \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})} + \sup_{\|\mathbf{a}\|_2=1}\sqrt{\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\right)\mathrm{Var}\left(\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\right)} \,. \end{aligned}$$

Thanks to Lemma 3, and Corollary 1 we have for any $\mathbf{a} \in \mathrm{S}^p$,

$$\begin{aligned}\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\right) &\leqslant \frac{2}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\right) + 2\kappa_n^4\mathrm{Var}\left(\mathrm{R}_{--}(\mathbf{D})\right) \\ &\leqslant \frac{2}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\right) + \frac{2\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n} \,. \end{aligned}$$

Furthermore, we have,

$$\begin{aligned}\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\right) &= \mathrm{Var}\left(\sum_{i,j=1}^{2}\left(\hat{\mathrm{B}}_{-2}(\mathbf{D})\right)_{i,j}\mathbf{a}^\top\Phi_2^{(i)}\Phi_2^{(j)\top}\right) \\ &\leqslant 4\sum_{i,j=1}^{2}\mathrm{Var}\left(\left(\hat{\mathrm{B}}_{-2}(\mathbf{D})\right)_{i,j}\mathbf{a}^\top\Phi_2^{(i)}\Phi_2^{(j)\top}\right) \\ &\leqslant \mathrm{c}\kappa_n^4 + \mathrm{c}\sum_{i,j=1}^{2}\mathrm{Var}\left(\mathbf{a}^\top\Phi_2^{(i)}\Phi_2^{(j)\top}\right) \leqslant \mathrm{c}\kappa_n^4 \,, \end{aligned}$$

similarly,

$$\mathrm{Var}\left(\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top}\right) \leqslant \mathrm{c}\kappa_n^4 \,. \tag{77}$$

Putting this all together, we have,

$$\left\|\mathbb{E}\left[\chi_1(\mathbf{D})\right]\right\|_2 \leqslant \frac{\mathrm{c}\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})} + \frac{\mathrm{c}\kappa_n^5\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}\sqrt{n}} \,. \tag{78}$$

Respectively for the variances of $\chi_1(\mathbf{D})$ and $\chi_2(\mathbf{D})$, we apply the law of total variance, to this end denote $M^{(i,j)}(\mathbf{D}) = \mathbb{E}[\Phi_1^{(1,2)\top}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(i,j)} \mid \mathcal{F}_{--}]$, first dealing with $\chi_2(\mathbf{D})$, we have

$$\begin{aligned}\mathrm{Var}\left(\chi_2(\mathbf{D})\right) &= \mathbb{E}\left[\mathrm{Var}\left(\chi_2(\mathbf{D}) \mid \mathcal{F}_{--}\right)\right] + \mathrm{Var}\left(M^{(1,2)}(\mathbf{D})\mathrm{R}_{--}(\mathbf{D})M^{(3,4)}(\mathbf{D})\right) \\ &\leqslant \mathbb{E}\left[\mathrm{Var}\left(\chi_2(\mathbf{D}) \mid \mathcal{F}_{--}\right)\right] + \mathbb{E}\left[\left\|M^{(1,2)}(\mathbf{D})\mathrm{R}_{--}(\mathbf{D})M^{(3,4)}(\mathbf{D})\right\|_2^2\right] \\ &\leqslant \mathbb{E}\left[\mathrm{Var}\left(\chi_2(\mathbf{D}) \mid \mathcal{F}_{--}\right)\right] + \frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^2} \,, \end{aligned}$$

and, from the independance of $(\Phi_1^{(1,2)}, \Phi_1^{(3,4)})$ with $(\Phi_2^{(1,2)}, \Phi_2^{(3,4)})$

$$
\begin{aligned}
&\mathrm{Var}\left(\chi_2(\mathbf{D}) \mid \mathcal{F}_{--}\right) \\
&\leqslant \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\Bigg\{ \mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right) \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right) \\
&\qquad + \left\|\mathbb{E}\left[\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right]\right\|_{\mathrm{op}}^2 \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)} \mid \mathcal{F}_{--}\right) \\
&\qquad + \mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right) \left\|\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\right\|_2^2 \Bigg\} \\
&\leqslant \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\Bigg\{ \mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right) \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right) \\
&\qquad + \kappa_n^4 \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right) \\
&\qquad + \kappa_n^4 \mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right) \Bigg\} ,
\end{aligned}
$$

similarly as for (77), it holds,

$$
\mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right) \leqslant cp\kappa_n^4 , \quad \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right) \leqslant c\kappa_n^4 ,
$$

thus,

$$
\mathrm{Var}\left(\chi_2(\mathbf{D})\right) \leqslant \frac{c\kappa_n^8 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2} . \tag{79}
$$

Finally, to control the variance of $\chi_1(\mathbf{D})$, first use the law of total variance to write,

$$
\begin{aligned}
\mathrm{Var}\left(\chi_1(\mathbf{D})\right) = \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathbb{E}\left[\left\|\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\right\|_{\mathrm{op}}^2 \mathrm{Var}\left(\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top} \mid \mathcal{F}_{-}\right)\right] \\
+ \mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\mathbb{E}[\Phi_1^{(1,2)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(3,4)\top} \mid \mathcal{F}_{-}]\right) ,
\end{aligned}
$$

proceeding similarly to (77) and using $\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top} \preceq \Phi_2^{(1,2)}\Phi_2^{(1,2)\top}$ on the first term, as well as Lemma 3 on the second yields,

$$
\begin{aligned}
\mathrm{Var}\left(\chi_1(\mathbf{D})\right) \leqslant \frac{c\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathbb{E}\left[\|\Phi_2^{(1,2)}\Phi_2^{(1,2)\top}\|_{\mathrm{op}}^2\right] + \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(1,2)\top}\right) \\
+ \kappa_n^4 \mathrm{Var}\left(\mathrm{R}_{--}(\mathbf{D})\mathbb{E}[\Phi_1^{(3,4)}\hat{\mathrm{B}}(\mathbf{D})\Phi_1^{(1,2)\top} \mid \mathcal{F}_{-}]\right) ,
\end{aligned}
$$

we conclude using that $\|xx^\top\|_{\mathrm{op}}^2 = \|x\|_2^4$,

$$
\mathrm{Var}\left(\chi_1(\mathbf{D})\right) \leqslant c\frac{\kappa_n^8 p^2}{\lambda_{\min}^{\Re}(\mathbf{D})^2} . \tag{80}
$$

Plugging the previous upper bounds (76), (78), (79) and (80) inside of (75), we get,

$$
\begin{aligned}
\left|s_2 - \tilde{s}_2^1\right| \leqslant c\Bigg(\frac{\sqrt{p}}{n^{5/2}}\left(\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^3} + \frac{\kappa_n^9}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}}\right) + \frac{\kappa_n^8 p^{3/2}}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n^2} \\
+ \frac{p}{n^{5/2}}\left(\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^2} + \frac{\kappa_n^9}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2}} + \frac{\kappa_n^9}{\lambda_{\min}^{\Re}(\mathbf{D})^{7/2}}\right)\Bigg) .
\end{aligned} \tag{81}
$$

In order to further simplify the expression of the equivalent $\tilde{s}_2^{(1)}$, we define,

$$
\tilde{s}_2^2 = \frac{n-1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\right]\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right]\mathbb{E}\left[\Phi_2^{(1,2)}\widetilde{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right] ,
$$

and we have,

$$
\begin{aligned}
&\frac{n}{n-1}\left(\tilde{s}_2^2 - \tilde{s}_2^1\right) \\
&= \mathbb{E}\left[\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\widetilde{\mathrm{B}}_{-1}(\mathbf{D}) - \hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right]\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\right] \\
&\quad + \mathbb{E}\left[\mathbb{E}\left[\Phi_1^{(3,4)}\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\right]\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\left\{\widetilde{\mathrm{B}}_{-2}(\mathbf{D}) - \hat{\mathrm{B}}_{-2}(\mathbf{D})\right\}\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\right] .
\end{aligned}
$$

It results, from Lemma 1

$$\begin{aligned}\left|\tilde{s}_2^1-\tilde{s}_2^2\right| \leqslant & \left|\mathbb{E}\left[\mathbb{E}\left[\Phi_1^{(3,4)}\left\{\widetilde{\mathrm{B}}_{-1}(\mathbf{D})-\hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(1,2)\top} \mid \mathcal{F}_{--}\right]\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\right]\right| \\ & + \kappa_n^2\left\|\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)}\left\{\widetilde{\mathrm{B}}_{-1}(\mathbf{D})-\hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\,\middle|\,\mathcal{F}_{--}\right]\right]\right\|_2 .\end{aligned}$$

Recall that for any vector $\mathbf{v}\in\mathbb{R}^p$, it holds, $\|\mathbf{v}\|_2=\sup_{\|\mathbf{a}\|_2=1}\mathbf{a}^\top\mathbf{v}$, thus using the Cauchy-Schwarz inequality and (45) for the difference $\widetilde{\mathrm{B}}_{-1}(\mathbf{D})-\hat{\mathrm{B}}_{-1}(\mathbf{D})$,

$$\begin{aligned}\left|\tilde{s}_2^1-\tilde{s}_2^2\right| \leqslant & \left|\mathbb{E}\left[\mathrm{tr}\left(\left\{\widetilde{\mathrm{B}}_{-1}(\mathbf{D})-\hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\Phi_1^{(3,4)}\right)\right]\right| \\ & + \kappa_n^2\sup_{\|\mathbf{a}\|_2=1}\mathbb{E}\left[\mathrm{tr}\left(\left\{\widetilde{\mathrm{B}}_{-1}(\mathbf{D})-\hat{\mathrm{B}}_{-1}(\mathbf{D})\right\}\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right)\right] \\ \leqslant & \frac{1}{n}\sqrt{\mathrm{Var}\left(\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right)\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\Phi_1^{(3,4)}\right)} \\ & + \sup_{\|\mathbf{a}\|_2=1}\frac{\kappa_n^2}{n}\sqrt{\mathrm{Var}\left(\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right)\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right)} .\end{aligned} \tag{82}$$

Towards bounding the previous error term, we write thanks to Corollary 1 and Lemma 3 that,

$$\mathrm{Var}\left(\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right) \leqslant \mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right) \leqslant \mathrm{c}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2}+\frac{\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right) ,$$

$$\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\right) \leqslant \frac{1}{n^2}\mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right) \leqslant \frac{\mathrm{c}}{n^2}\left(\frac{\kappa_n^4 p}{\lambda_{\min}^{\Re}(\mathbf{D})^2}+\frac{\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right) ,$$

$$\begin{aligned}\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right) & \leqslant \mathrm{Var}\left(\Phi_1^{(3,4)\top}\mathbf{a}^\top\mathrm{R}_{--}(\mathbf{D})\Phi_1^{(1,2)}\right)+\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\right) \\ & \leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})^2}+\frac{\kappa_n^6 p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right)+\frac{\mathrm{c}p}{n^2}\left(\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^4}+\frac{\kappa_n^{10}}{\lambda_{\min}^{\Re}(\mathbf{D})^5 n}\right) .\end{aligned}$$

and,

$$\begin{aligned} & \mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\Phi_1^{(3,4)}\right) \\ \leqslant & \mathrm{Var}\left(\Phi_1^{(1,2)\top}\mathrm{R}_{--}(\mathbf{D})\mathbb{E}\left[\Phi_2^{(1,2)}\hat{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top} \mid \mathcal{F}_{--}\right]\Phi_1^{(3,4)}\right)+\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^2}\mathrm{Var}\left(\hat{\mathrm{B}}_{-1}(\mathbf{D})\right) \\ \leqslant & \mathrm{c}\left(\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^2}+\frac{\kappa_n^{10} p}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n}\right)+\frac{\mathrm{c}p}{n^2}\left(\frac{\kappa_n^{12}}{\lambda_{\min}^{\Re}(\mathbf{D})^4}+\frac{\kappa_n^{14}}{\lambda_{\min}^{\Re}(\mathbf{D})^5 n}\right) .\end{aligned}$$

Plugging the previous fours variance bounds in (82), we conclude

$$\begin{aligned}\left|\tilde{s}_2^1-\tilde{s}_2^2\right| \leqslant & \frac{\mathrm{c}\left(1+\kappa_n^4\right)\sqrt{p}}{n}\left(\frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})}+\frac{\kappa_n^3}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}\sqrt{n}}\right) \\ & \times\left(\frac{\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})}+\frac{\kappa_n^3\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}\sqrt{n}}+\frac{\kappa_n^4\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n}+\frac{\kappa_n^5\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2}n^{3/2}}\right) .\end{aligned} \tag{83}$$

Finally, we show that $\tilde{s}_2^2$ is close to $\overline{s}_2$ defined as

$$\overline{s}_2=\mathbb{E}\left[\Phi_1^{(3,4)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)\top}\right]\overline{\mathrm{R}}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)}\mathrm{B}(\mathbf{D})\Phi_1^{(1,2)}\right] ,$$

recall

$$\tilde{s}_2^2=\frac{n-1}{n}\mathbb{E}\left[\Phi_1^{(3,4)}\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\Phi_1^{(1,2)\top}\right]\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right]\mathbb{E}\left[\Phi_2^{(1,2)}\widetilde{\mathrm{B}}_{-2}(\mathbf{D})\Phi_2^{(3,4)\top}\right] ,$$

then it holds

$$\left|\widetilde{s}_2^2 - \overline{s}_2\right| \leqslant \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})n} + \frac{2\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})}\left\|\mathrm{B}(\mathbf{D}) - \widetilde{\mathrm{B}}_{-1}(\mathbf{D})\right\|_{\mathrm{op}} + \kappa_n^4\left\|\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} ,$$

where we have used the i.i.d porperty of $(\Phi_1^{(1,2)}, \Phi_1^{(3,4)})$ and $(\Phi_2^{(1,2)}, \Phi_2^{(3,4)})$, we conclude on this error term by using (45) which gives,

$$\begin{aligned}\left\|\mathrm{B}(\mathbf{D}) - \widetilde{\mathrm{B}}_{-1}(\mathbf{D})\right\|_{\mathrm{op}} &= \frac{1}{n}\left\|\mathrm{B}(\mathbf{D})\mathbb{E}\left[\Phi_1^{(1,2)\top}\mathbb{E}\left[\mathrm{R}(\mathbf{D}) - \mathrm{R}_{--}(\mathbf{D})\right]\Phi_1^{(1,2)}\right]\widetilde{\mathrm{B}}_{-1}(\mathbf{D})\right\|_{\mathrm{op}} \\ &\leqslant \frac{\kappa_n^2}{n}\left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D}) - \mathrm{R}_{--}(\mathbf{D})\right]\right\|_{\mathrm{op}} \leqslant \frac{2\kappa_n^2}{\lambda_{\min}^{\Re}(\mathbf{D})n} .\end{aligned}$$

It results,

$$\begin{aligned}\left|\widetilde{s}_2^2 - \overline{s}_2\right| &\leqslant \frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})n} + \frac{4\kappa_n^6}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n} + \kappa_n^2\left\|\mathbb{E}\left[\mathrm{R}_{--}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} \\ &\leqslant \mathrm{c}\left(\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})n} + \frac{\kappa_n^6}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n} + \kappa_n^2\left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}}\right) .\end{aligned} \tag{84}$$

Putting this all together, we use the triangle inequality to get,

$$\begin{aligned}|s_2 - \overline{s}_2| \leqslant\ & \mathrm{c}\kappa_n^4\left\|\mathbb{E}\left[\mathrm{R}(\mathbf{D})\right] - \overline{\mathrm{R}}(\mathbf{D})\right\|_{\mathrm{F}} + \frac{\mathrm{c}(1+\kappa_n^{10})\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n}\left(1 + \frac{\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}\sqrt{n}}\right) + \frac{cp^{3/2}}{n^2}\frac{\kappa_n^8}{\lambda_{\min}^{\Re}(\mathbf{D})^3} \\ &+ \frac{\mathrm{c}(1+\kappa_n^{10})\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{5/2} n^{3/2}}\left(1 + \frac{\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}\sqrt{n}}\right) + \frac{\mathrm{c}}{n}\frac{\kappa_n^4}{\lambda_{\min}^{\Re}(\mathbf{D})} \\ &+ \frac{\mathrm{c}(1+\kappa_n^{10})p}{\lambda_{\min}^{\Re}(\mathbf{D})^2 n^{5/2}}\left(1 + \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}} + \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^{3/2}}\right) \\ &+ \frac{cp}{n^3}\frac{1+\kappa_n^{10}}{\lambda_{\min}^{\Re}(\mathbf{D})^4} + \frac{\mathrm{c}(1+\kappa_n^{10})\sqrt{p}}{\lambda_{\min}^{\Re}(\mathbf{D})^3 n^{5/2}}\left(1 + \frac{1}{\lambda_{\min}^{\Re}(\mathbf{D})^{1/2}}\right) .\end{aligned} \tag{85}$$

Where the last inequality follows from (81), (83) and (84). The statement follows from (74) and (85) and the triangle inequality. □

Similarly as for Proposition 6, Proposition 7 may be directly applied to approximate $\mathbb{E}\left[\xi_{\alpha,\lambda}(\zeta)\right]$ defined in (35), for all $\zeta \in \mathsf{U}_\lambda$. Indeed, setting $\mathbf{D} = \alpha\mathbb{E}\left[\Lambda(\mathtt{Z})\right] + \lambda_n \mathrm{I}_p + \zeta\Sigma$ and $\Psi = (\Psi_i)_{i=1}^n$ s.t. for all $i \leqslant n$, $\Psi_i = (\sqrt{1-\alpha}\varphi(\mathtt{X}_i), \sqrt{\alpha}\mu_x(\mathtt{Z}_i), \sqrt{1-\alpha}\mathtt{Y}_i, \sqrt{\alpha}\mu_y(\mathtt{Z}_i))$, yields $\Xi(\mathbf{D}) = \xi_{\alpha,\lambda}(\zeta)$. Thanks to our assumptions on $\mathtt{Z}$ and $\mathcal{T}$, in **H**1 and **H**4, it is clear that such Ψ and $\mathbf{D}$ satisfy the assumptions of Proposition 7, which ensure $\mathbb{E}\left[\xi_{\alpha,\lambda}(\zeta)\right] = \overline{\Xi}(\mathbf{D})$.

In addition to Lemma 9, this results allows to approxmiate $\mathbb{E}\left[\hat{\theta}_{\alpha,\lambda}^\top \Sigma \hat{\theta}_{\alpha,\lambda}\right]$, which will be detailed in Section F.

F Proofs of Proposition 1, and Theorem 1

This final section is dedicated to the proofs of the main results of the paper, namely Proposition 1, and Theorem 1.

We begin by writting the proof of Proposition 1 in full detail, as it will be used to prove Theorem 1. To this end, we leverage Theorem 2 as well as Proposition 6, Proposition 7 and Lemma 9:

Proof of Proposition 1. We will use the following notation, for any $\zeta_n \in \mathsf{U}_{\lambda_n}$

$$\begin{aligned}\mathbf{D}_{\alpha,\lambda_n}(\zeta_n) &= \alpha\mathbb{E}\left[\Lambda(\mathtt{Z})\right] + \lambda_n \mathrm{I}_p + \zeta_n\Sigma , \\ \widetilde{\mathrm{R}}_{\alpha,\lambda_n}(\zeta_n) &= \left((1-\alpha)\mathrm{C} + \alpha\mathrm{C}' + \mathbf{D}_{\alpha,\lambda_n}(\zeta_n)\right)^{-1} , \\ \widetilde{\mathrm{H}}_\alpha &= (1-\alpha)\mathrm{H} + \alpha\mathrm{H}' + \alpha\mathbb{E}\left[\Omega(\mathtt{Z})\right] , \\ \widetilde{\theta}_{\alpha,\lambda_n} &= \widetilde{\mathrm{R}}_{\alpha,\lambda_n}(0)\widetilde{\mathrm{H}}_\alpha , \\ \widetilde{\xi}_{\alpha,\lambda_n}(\zeta_n) &= \widetilde{\mathrm{H}}_\alpha \widetilde{\mathrm{R}}_{\alpha,\lambda_n}(\zeta_n)\widetilde{\mathrm{H}}_\alpha .\end{aligned}$$

As well as,

$$\forall i \leqslant n\ , \quad \Phi_i = (\varphi(\mathtt{X}_i), \mu_x(\mathtt{Z}_i))\ , \qquad \Psi_i = (\mathtt{Y}_i^\top, \mu_y(\mathtt{Z}_i)^\top)^\top\ ,$$

$$\mathrm{W}_\alpha = \mathrm{Diag}((\sqrt{\alpha}, \sqrt{1-\alpha}))\ , \qquad \mathrm{B}_{\alpha,\lambda_n}(\zeta_n) = \mathrm{W}_\alpha \Big(\mathrm{I}_2 + n^{-1} \mathrm{W}_\alpha \mathbb{E}\left[\Phi_1^\top \mathbb{E}\left[\mathrm{R}_{\alpha,\lambda_n}(\zeta_n) \right] \Phi_1 \right] \mathrm{W}_\alpha^\top \Big)^{-1} \mathrm{W}_\alpha\ ,$$

$$\overline{\Sigma}(\zeta_n) = \mathbb{E}\big[\Phi_1 \mathrm{B}_{\alpha,\lambda_n}(\zeta_n) \Phi_1^\top \big]\ , \qquad \overline{\Gamma}(\zeta_n) = \mathbb{E}\big[\Phi_1 \mathrm{B}_{\alpha,\lambda_n}(\zeta_n) \Psi_1^\top \big]\ ,$$

$$\overline{\mathrm{R}}_{\alpha,\lambda_n}(\zeta_n) = \big(\overline{\Sigma}(\zeta_n) + \mathbf{D}_{\alpha,\lambda_n}(\zeta_n) \big)^{-1}\ , \qquad \overline{\mathrm{H}}_\alpha(\zeta_n) = \overline{\Gamma}(\zeta_n) + \mathbb{E}\left[\Omega(\mathtt{Z}) \right]\ ,$$

$$\overline{\theta}_{\alpha,\lambda_n} = \overline{\mathrm{R}}_{\alpha,\lambda_n}(0) \overline{\mathrm{H}}_\alpha(0)\ , \qquad \overline{\xi}_{\alpha,\lambda_n}(\zeta_n) = \mathbb{E}\left[\Psi_1 \left\{ \mathrm{I}_2 - \mathrm{B}_{\alpha,\lambda_n}(\zeta_n) \right\} \Psi_1^\top \right] + \overline{\mathrm{H}}_\alpha(\zeta_n)^\top \overline{\mathrm{R}}_{\alpha,\lambda_n}(\zeta_n) \overline{\mathrm{H}}_\alpha(\zeta_n)\ .$$

Note that the previous definition either match the ones of Proposition 1, or generalize them to the case of a complex-value regularization $\zeta_n \Sigma$ in the resolvent. We have thanks to the variance bounds on Λ and Ω in **H**4, we have,

$$\mathbb{E}\left[\| \Lambda(\mathtt{Z}) - \mathbb{E}\left[\Lambda(\mathtt{Z}) \right] \|_{\mathrm{F}}^2 \right] \leqslant \frac{\mathrm{c}}{n}\ , \quad \mathbb{E}\left[\| \Omega(\mathtt{Z}) - \mathbb{E}\left[\Omega(\mathtt{Z}) \right] \|_2^2 \right] \leqslant \frac{\mathrm{c}}{n}\ ,$$

hence for any $\zeta_n \in \mathsf{U}_{\lambda_n}$, we have,

$$\begin{aligned} \left\| \mathbb{E}\left[\mathrm{R}_{\alpha,\lambda_n} \right] - \mathbb{E}\left[\widetilde{\mathrm{R}}_{\alpha,\lambda_n}(0) \right] \right\|_{\mathrm{F}} &\leqslant \frac{\mathrm{c}}{\lambda_n^2 \sqrt{n}}\ , \\ \left\| \mathbb{E}\left[\hat{\theta}_{\alpha,\lambda_n} \right] - \mathbb{E}\left[\widetilde{\theta}_{\alpha,\lambda_n} \right] \right\|_2 &\leqslant \frac{\mathrm{c}}{\lambda_n^2 \sqrt{n}}\ , \\ \left| \mathbb{E}\left[\xi_{\alpha,\lambda_n}(\zeta_n) \right] - \mathbb{E}\left[\widetilde{\xi}_{\alpha,\lambda_n}(\zeta_n) \right] \right| &\leqslant \frac{\mathrm{c}}{\lambda_n^2 \sqrt{n}}\ . \end{aligned} \tag{86}$$

We prove each point of Proposition 1 separately, to this end, we leverage each bound provided in (86).

1. *Deterministic equivalent for* $\mathrm{R}_{\alpha,\lambda_n}$.

   First recall from the proof of Proposition 4, that $\mathrm{R}_{\alpha,\lambda_n}$ writes as,

   $$\mathrm{R}_{\alpha,\lambda_n} = r(\mathtt{Z}, 0)\ ,$$

   where r was defined in (40). Furthermore, we have shown in (41) that $r(\cdot, 0)$ is $\mathtt{L}^r$-Lipschitz continuous, with

   $$\mathtt{L}_{\mathtt{K}_n}^r \leqslant \frac{\mathrm{c}(1+\sqrt{\lambda_n})}{\lambda_n^{3/2}\sqrt{n}} \leqslant \frac{\mathrm{c}}{\lambda_n^{3/2}\sqrt{n}}\ ,$$

   where we have used the hypothesis $\lambda_n \geqslant 1/n$ to simplify the bound on the Lipschitz constant. It results that the function $\mathbf{Z} \mapsto \mathrm{tr}\left(\mathbf{A} r(\mathbf{Z}, 0) \right)$ is $\mathrm{c}\lambda_n^{-3/2} n^{-1/2}$-Lipschitz continuous, hence the Lipschitz concentration hypothesis on $\mathtt{Z}$ **H** 1 yields,

   $$\left| \mathrm{tr}\left(\mathbf{A} r(\mathtt{Z}, 0) \right) - \mathbb{E}\left[\mathrm{tr}\left(\mathbf{A} r(\mathtt{Z}, 0) \right) \right] \right| \prec_s \frac{\mathrm{c}}{\lambda_n^{3/2}\sqrt{n}}\ . \tag{87}$$

   Furthermore, using Proposition 5 with $\Psi_i \leftarrow (\sqrt{1-\alpha}\varphi(\mathtt{Z}_i), \sqrt{\alpha}\mu_x(\mathtt{Z}_i))$, and $\mathbf{D} \leftarrow \mathbf{D}_{\alpha,\lambda_n}(0)$, we have,

   $$\left\| \mathbb{E}\left[\widetilde{\mathrm{R}}_{\alpha,\lambda_n} \right] - \overline{\mathrm{R}}_{\alpha,\lambda_n} \right\|_{\mathrm{F}} \leqslant \frac{\mathrm{c}}{\lambda_n^{7/2}\sqrt{n}}\ , \tag{88}$$

   where we have used $\kappa < +\infty$ and $\limsup p/n < +\infty$ granted by **H**2 to simplify the bound. Combining (86), (87) and (88) concludes the proof of the first point.

2. *Deterministic equivalent for* $\hat{\theta}_{\alpha,\lambda_n}$.

   Recall from Proposition 4 that, for any sequence $(\mathbf{a}_n)_{n\in\mathbb{N}}$ of vectors in $\mathbb{R}^p$,

   $$\left| \mathbf{a}_n^\top \hat{\theta}_{\alpha,\lambda_n} - \mathbb{E}\left[\mathbf{a}_n^\top \hat{\theta}_{\alpha,\lambda_n} \right] \right| \prec_s \frac{\|\mathbf{a}_n\|_2}{\lambda_n^{3/2}\sqrt{n}} \left(1 + \sqrt{\lambda_n} \right) \leqslant \frac{\mathrm{c}\|\mathbf{a}_n\|_2}{\lambda_n^{3/2}\sqrt{n}}\ , \tag{89}$$

which we have simplified thanks to $\lambda_n \geqslant 1/n$. Furthermore, calling Proposition 6 with $\Psi_i \leftarrow (\sqrt{1-\alpha}\varphi(\mathtt{Z}_i)^\top, \sqrt{\alpha}\mu_x(\mathtt{Z}_i)^\top, \sqrt{1-\alpha}\mathtt{Y}_i^\top, \sqrt{\alpha}\mu_y(\mathtt{Z}_i)^\top)^\top$, and $\mathbf{D} \leftarrow \mathbf{D}_{\alpha,\lambda_n}(0)$, directly yields,

$$\left\| \mathbb{E}\left[\widetilde{\theta}_{\alpha,\lambda_n}\right] - \overline{\theta}_{\alpha,\lambda_n} \right\|_2 \leqslant \frac{\mathrm{c}}{\lambda_n^4\sqrt{n}} , \tag{90}$$

where the bound of Proposition 6 has been simplified thanks to $\lambda_n \geqslant 1/n$, $\kappa < +\infty$ and $\limsup p/n < +\infty$ granted by **H**2. Combining (86), (89) and (90) concludes the proof of the second point.

3. *Deterministic equivalent for* $\xi_{\alpha,\lambda_n}(r_n \mathrm{e}^{\mathrm{i}t})$.

Recall from Proposition 4 that, for any sequence $(r_n)_{n\in\mathbb{N}}$ of real numbers,

$$\left|\xi_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t}) - \mathbb{E}\left[\xi_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})\right]\right| \prec_s \frac{1+r_n}{(\lambda_n - r_n\mathrm{L}_\varphi^2)^{3/2}\sqrt{n}}\left(1+\sqrt{\lambda_n - r_n\mathrm{L}_\varphi^2}\right) , \tag{91}$$

in particular, for $r_n = 0$, and thanks to $\lambda_n \geqslant 1/n$, we have,

$$\left|\xi_{\alpha,\lambda_n}(0) - \mathbb{E}\left[\xi_{\alpha,\lambda_n}(0)\right]\right| \prec_s \frac{\mathrm{c}}{\lambda_n^{3/2}\sqrt{n}} . \tag{92}$$

Furthermore, calling Proposition 7 with $\Psi_i \leftarrow (\sqrt{1-\alpha}\varphi(\mathtt{Z})^\top, \sqrt{\alpha}\mu_x(\mathtt{Z})^\top, \sqrt{1-\alpha}\mathtt{Y}^\top, \sqrt{\alpha}\mu_y(\mathtt{Z})^\top)^\top$, and $\mathbf{D} \leftarrow \mathbf{D}_{\alpha,\lambda_n}(0)$, we have,

$$\left|\mathbb{E}\left[\widetilde{\xi}_{\alpha,\lambda_n}(0)\right] - \overline{\xi}_{\alpha,\lambda_n}(0)\right| \leqslant \frac{\mathrm{c}}{\lambda_n^{7/2}\sqrt{n}} , \tag{93}$$

where the bound of Proposition 7 has been simplified thanks to $\lambda_n \geqslant 1/n$, $\kappa < +\infty$ and $\limsup p/n < +\infty$ granted by **H**2. Combining (86), (92) and (93) concludes the proof of the third point.

4. *Deterministic equivalent for* $\hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n}$.

First leverage (91) as well as Lemma 9 to show that for any sequence $(r_n)_{n\in\mathbb{N}}$ of positive numbers,

$$\left|\partial_\zeta \xi_{\alpha,\lambda_n}(0) - \mathbb{E}\left[\partial_\zeta \xi_{\alpha,\lambda_n}(0)\right]\right| \prec_s \frac{r_n^{-1}+1}{(\lambda_n - r_n\mathrm{L}_\varphi^2)^{3/2}\sqrt{n}}\left(1+\sqrt{\lambda_n - r_n\mathrm{L}_\varphi^2}\right) ,$$

taking $r_n = \lambda_n/(2\mathrm{L}_\varphi^2)$, and recalling that $\partial_\zeta \xi_{\alpha,\lambda_n}(0) = -\hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n}$, we have,

$$\left|\hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n} + \mathbb{E}\left[\partial_\zeta \xi_{\alpha,\lambda_n}(0)\right]\right| \prec_s \frac{1}{\lambda_n^{5/2}\sqrt{n}} . \tag{94}$$

In addition, applying Lemma 9 to the third equation in (86), we have,

$$\sup_{t\in[0,2\pi]} \left|\mathbb{E}\left[\xi_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})\right] - \mathbb{E}\left[\widetilde{\xi}_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})\right]\right| \leqslant \frac{\mathrm{c}}{\lambda_n^2\sqrt{n}} ,$$

thus,

$$\left|\mathbb{E}\left[\partial_\zeta \xi_{\alpha,\lambda_n}(0)\right] - \mathbb{E}\left[\partial_\zeta \widetilde{\xi}_{\alpha,\lambda_n}(0)\right]\right| \leqslant \frac{\mathrm{c}}{\lambda_n^3\sqrt{n}} . \tag{95}$$

Finally, from Proposition 7 applied with $\Psi_i \leftarrow (\sqrt{1-\alpha}\varphi(\mathtt{Z}_i)^\top, \sqrt{\alpha}\mu_x(\mathtt{Z}_i)^\top, \sqrt{1-\alpha}\mathtt{Y}_i^\top, \sqrt{\alpha}\mu_y(\mathtt{Z}_i)^\top)^\top$, and $\mathbf{D} \leftarrow \mathbf{D}_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})$, we have,

$$\sup_{t\in[0,2\pi]} \left|\mathbb{E}\left[\widetilde{\xi}_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})\right] - \overline{\xi}_{\alpha,\lambda_n}(r_n\mathrm{e}^{\mathrm{i}t})\right| \leqslant \frac{\mathrm{c}}{\lambda_n^{7/2}\sqrt{n}} ,$$

thus Lemma 9 yields,

$$\left|\mathbb{E}\left[\partial_\zeta \widetilde{\xi}_{\alpha,\lambda_n}(0)\right] - \partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0)\right| \leqslant \frac{\mathrm{c}}{\lambda_n^{9/2}\sqrt{n}} . \tag{96}$$

Combining (94), (95) and (96) we conclude:

$$\left|\hat{\theta}_{\alpha,\lambda_n}^\top \Sigma \hat{\theta}_{\alpha,\lambda_n} + \partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0)\right| \prec_s \frac{1}{\lambda_n^{9/2}\sqrt{n}} \, .$$

Finally, we explicitely compute the diferential $\partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0)$ as follows:

$$\partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0) = -\mathbb{E}\left[\Psi_1 \partial_\zeta \mathrm{B}_{\alpha,\lambda_n}(0) \Psi_1^\top\right] - \overline{\mathrm{H}}_\alpha(0)^\top \mathrm{R}_{\alpha,\lambda_n}(0) \partial_\zeta \overline{\Sigma}(0) \mathrm{R}_{\alpha,\lambda_n}(0) \overline{\mathrm{H}}_\alpha(0) + 2(\partial_\zeta \overline{\mathrm{H}}_\alpha(0))^\top \mathrm{R}_{\alpha,\lambda_n}(0) \overline{\mathrm{H}}_\alpha(0) \, ,$$

and verify that,

$$\partial_\zeta \mathrm{B}_{\alpha,\lambda_n}(0) = \mathrm{D}_{\alpha,\lambda_n} \, , \qquad \partial_\zeta \overline{\Sigma}(0) = \mathbb{E}\left[\Phi_1 \mathrm{D}_{\alpha,\lambda} \Phi_1\right] + \Sigma \, , \qquad \partial_\zeta \overline{\mathrm{H}}_\alpha(0) = \mathbb{E}\left[\Phi_1 \mathrm{D}_{\alpha,\lambda} \Psi_1\right] \, ,$$

where $\mathrm{D}_{\alpha,\lambda_n}$ is the matrix defined in (12). This concludes the proof of Proposition 1.

□

Lastly, we prove Theorem 1,

Proof of Theorem 1. We begin by recalling the definition of $\overline{\mathcal{G}}(\alpha, \lambda_n)$, we have,

$$\overline{\mathcal{G}}(\alpha, \lambda_n) = \theta_\star^\top \Sigma_{\star\star} \theta_\star + \sigma^2 - 2\theta_\star^\top \Sigma_\star \overline{\theta}_{\alpha,\lambda_n} + \overline{\theta}_{\alpha,\lambda_n}^\top \{\overline{\Sigma}' + \Sigma\} \overline{\theta}_{\alpha,\lambda_n} - 2\overline{\theta}_{\alpha,\lambda_n}^\top \overline{\Gamma}' + \overline{\gamma} \, ,$$

where one might identify in the previous term $-\partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0) = \overline{\theta}_{\alpha,\lambda_n}^\top \{\overline{\Sigma}' + \Sigma\} \overline{\theta}_{\alpha,\lambda_n} - 2\overline{\theta}_{\alpha,\lambda_n}^\top \overline{\Gamma}' + \overline{\gamma}$.
In addition, we have from (36) that,

$$\mathcal{G}(\alpha, \lambda_n) = \theta_\star^\top \Sigma_{\star\star} \theta_\star + \sigma^2 - 2\theta_\star^\top \Sigma_\star \hat{\theta}_{\alpha,\lambda_n} - \partial_\zeta \xi_{\alpha,\lambda_n}(0) \, ,$$

thus,

$$\left|\mathcal{G}(\alpha, \lambda_n) - \overline{\mathcal{G}}(\alpha, \lambda_n)\right| \leqslant 2\left|\theta_\star^\top \Sigma_\star \left\{\hat{\theta}_{\alpha,\lambda_n} - \overline{\theta}_{\alpha,\lambda_n}\right\}\right| + \left|\partial_\zeta \xi_{\alpha,\lambda_n}(0) - \partial_\zeta \overline{\xi}_{\alpha,\lambda_n}(0)\right| \, .$$

The result follows directly from Proposition 1. □