Universality of High-Dimensional Logistic Regression and a Novel CGMT under Dependence with Applications to Data Augmentation

Matthew Esmaili Mallory MATTHEWMALLORY@FAS.HARVARD.EDU
Department of Statistics, Harvard University
Kevin Han Huang HAN.HUANG.20@UCL.AC.UK
Gatsby Unit, University College London
Morgane Austern MAUSTERN@FAS.HARVARD.EDU
Department of Statistics, Harvard University

Editors: Nika Haghtalab and Ankur Moitra

Abstract

Over the last decade, a wave of research has characterized the exact asymptotic risk of many high-dimensional models in the proportional regime. Two foundational results have driven this progress: Gaussian universality, which shows that the asymptotic risk of estimators trained on non-Gaussian and Gaussian data is equivalent, and the convex Gaussian min-max theorem (CGMT), which characterizes the risk under Gaussian settings. However, these results rely on the assumption that the data consists of independent random vectors—an assumption that significantly limit its applicability to many practical setups. In this paper, we address this limitation by generalizing both results to the dependent setting. More precisely, we prove that Gaussian universality still holds for high-dimensional logistic regression under block dependence, m-dependence and special cases of β-mixing, and establish a novel CGMT framework that accommodates for correlation across both the covariates and observations. Using these results, we establish the impact of data augmentation, a widespread practice in deep learning, on the asymptotic risk.

Keywords: universality, logistic regression, high dimensions, CGMT, binary classification, block dependence, m-dependence, mixing, proportional asymptotics

1. Introduction

Over the past decade, landmark results such as Gaussian universality and the convex Gaussian min-max theorem (CGMT) have been extended and applied to analyze the asymptotic risk of various high-dimensional feature models. They have led to a deeper understanding of matters such as the impact of regularization and hyperparameters on the risk (Salehi et al., 2019; Deng et al., 2022) and the double descent phenomenon (Mei and Montanari, 2022; Hastie et al., 2022; Belkin et al., 2019).

Broadly speaking, Gaussian universality is the observation that the risk of many high dimensional estimators depends on the data distribution only through its first two moments (Montanari and Saeed, 2022; Montanari et al., 2023; Dandi et al., 2024; Gerace et al., 2024; Korada and Montanari, 2011; Han and Shen, 2023; Hu and Lu, 2022). Consequently, for these estimators, their risks can be studied by analyzing the risk for Gaussian data with matching mean and variance. This unlocks the many useful tools developed for the Gaussian case, including Approximate Message Passing (Donoho et al., 2009), the Cavity Method (Opper et al., 2001) and the CGMT (Gordon, 1985; Thrampoulidis et al., 2014). Among them, the CGMT is a framework that converts a complex optimization problem on Gaussian data to a much more analytically tractable auxiliary problem. The auxiliary optimization is often further simplified into a deterministic equation involving only a few scalars, and under the CGMT, its solution completely characterizes that of the original problem.

A general pipeline of analysis built on universality and the CGMT entails the following:

(i) Consider a high-dimensional model, such as generalized linear regression or random feature models, with data following some pre-specified distribution;
(ii) Equate our estimation problem to that of the same model on Gaussian data via universality;
(iii) Simplify the Gaussian optimization problem via the CGMT into a format that can be more readily solved, either analytically or computationally.

One substantial limitation of existing Gaussian universality and CGMT analyses is that the data must consist of independent—and often also identically distributed—vectors, which is not realistic for many applications. Several forms of dependence are commonly observed in practice:

- *Block dependence.* An important example of dependence in machine learning is found in data augmentation[1], a technique that synthetically expands a training dataset by applying random transformations to existing data and incorporating the transformed data back into the dataset (Taqi et al., 2018; Shorten et al., 2021; Volkova, 2024). In machine learning practice, data augmentation has become one of the most widely adopted methods, especially in the presence of invariance (e.g. symmetries) or an underlying structure (e.g sparsity) (Lyle et al., 2019). Theoretically, however, the dependence arising from multiple transformed copies of the same observation makes the effect of data augmentation challenging to analyze.
- *m-dependence.* Another common form of dependence manifests through a finite dependency neighborhood: Under spatial moving average models (Cressie, 1993), the observation at a given point is dependent on a local neighborhood of observations but no further. Similar examples are ubiquitous in time series, graph and spatial analysis (Cryer, 1986; Brock et al., 1992; Schweinberger and Handcock, 2015; Wackernagel, 2003);
- *β-mixing.* Data can also depend on infinitely many variables, with strong short-range dependence and decaying long-range dependence. This is typically described by mixing conditions (Billingsley, 1995; Bradley, 2005), and is also found in many common time series and spatial models (Deo, 1973; Tuan and Lanh, 1985; Tsay, 2005; Gelfand et al., 2010).

This paper, for the first time, extends the Gaussian universality principle beyond the independence assumption to encompass dependent vectors (X_i) in the context of high-dimensional logistic regression. Universality results are provided for block dependence, m-dependence, as well as specific β-mixing processes. Moreover, we develop a novel CGMT framework, accommodating dependence both between covariates and observations under a certain "low-rank" assumption. Leveraging these two new tools, we precisely characterize the impact of data augmentation on the risk. We notably investigate the effectiveness of data augmentation when the invariance or structure of the problem is only partially known, as is often the case in practice (Benton et al., 2020; Yang et al., 2023).

1.1. Model Overview

We observe high-dimensional data $(X_i, y_i)_{i=1}^n$ with covariates $X_i \in \mathbb{R}^p$ and labels $y_i \equiv y_i(X_i) \in \{0, 1\}$. We consider the *proportional regime*, where the signal dimension p grows linearly with the sample size n. In our main universality result (Section 3), which is used for analyzing data augmentation, the data $\mathbf{X} := (X_i)_{i\leq n}$ — not assumed to be identically distributed — are block dependent:

$$(X_i, y_i) \perp\!\!\!\perp (X_j, y_j) \;\text{ if }\; j \notin \mathcal{B}_i := \left\{k\lfloor \tfrac{i-1}{k} \rfloor + 1, \ldots, k\lfloor \tfrac{i-1}{k} \rfloor + k\right\}. \tag{1}$$

[1] The definition of data augmentation in machine learning differs from its use in statistics. In the latter, data augmentation often refers to the introduction of latent variables to the model, e.g. in the EM algorithm.

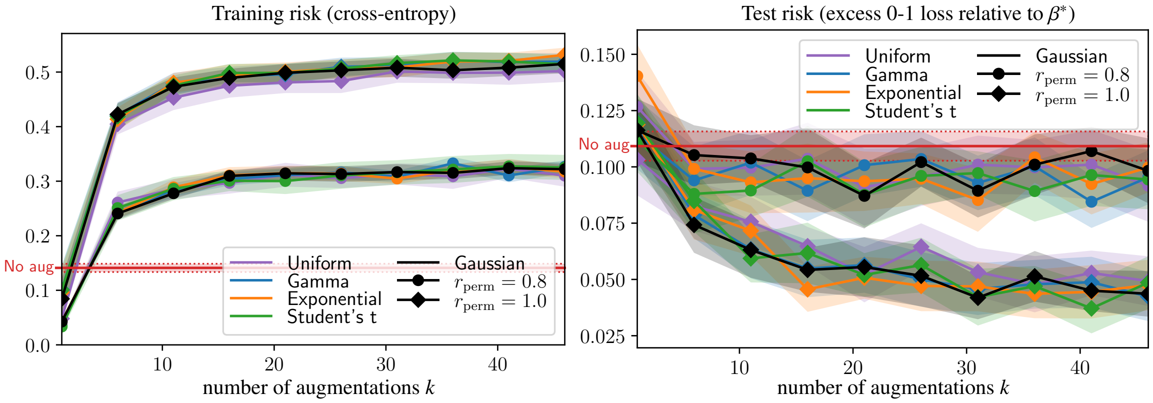

Figure 1: Universality of risks of a logistic regressor, trained with different number and amount of random permutations. See Section 6 and Appendix C for the detailed setup.

We also consider data that satisfy m-dependence and β-mixing; see Section 4 for the precise definitions. To relate the labels to their covariates, we assume there is a true signal $\beta^* \in \mathbb{R}^p$ such that

$$\mathbb{P}(y_i = 1 \mid \mathbf{X}) = \sigma(X_i^\top \beta^*), \qquad \sigma(t) := (1 + e^{-t})^{-1}. \tag{2}$$

The signal is estimated via a penalized and weighted logistic regression:

$$\hat{\beta}(\mathbf{X}) := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \omega_i \left(\log\left(1 + e^{X_i^\top \beta}\right) - y_i X_i^\top \beta \right) + \frac{\lambda}{2n} \|\beta\|^2, \tag{3}$$

where $(\omega_i) \in [0, 1]^{\mathbb{N}}$ are deterministic weights. If the weights are all set as 1, (3) recovers the traditional penalized logistic regression. More generally, ω_i's can be chosen to be different to accommodate potential heterogeneity such as heteroskedasticity (e.g., Shalizi (2019)), or unique cases like data augmentation (Section 6). Examples of setups that can be handled by this model include:

- *Dependent* (X_i) *and conditionally independent* (y_i)*:* Dependent covariates commonly arise in various applications. For instance, in biological experiments on mice, the littermate effect introduces dependence between the behaviors of mice from the same litter (Haseman and Kupper, 1979). Similarly, local dependence is prevalent in genomic data (Yu and Bien, 2017). In those settings, while the covariates are dependent, each response variable y_i has direct dependence only on X_i and not on the other covariates.

- *Block Dependent* (X_i) *and* (y_i). In many other practical settings, the response variable y_i can depend on a set of multiple covariates $\{X_j : j \in \mathcal{B}_i\}$. In ICU settings, for example, predicting 24-hour mortality is improved by incorporating past data on the same patient (Plate et al., 2019). To model these setups, we can assume that there exists a matrix $A \in \mathbb{R}^{n \times n}$ that models the dependencies between observations, so that y_i depends on the linear combination $Z_i := \sum_{j \in \mathcal{B}_i} a_{i,j} X_j$ of predictors in the same block (Wu and Ware, 1979). In this case, (Z_i, y_i) is still a block dependent process that satisfies all the assumptions of our setup.

- *Data Augmentation*. If the original dataset (Z_i) are independent, then the augmented data will exhibit block dependence within each set of augmented copies of X_i, whereas the response y_i typically depends only on the original variable Z_i; more details in Section 6.

Remark 1 *In* (3)*, the logistic regression is performed on the same variables that* y_i *depends on. We have chosen this presentation for simplicity. Appendix* B.3 *includes a more general model,* *which allows for the label to depend on the entire block $\{X_j : j \in \mathcal{B}_i\}$ while X_i is regressed only on a subset of those observations. This hence allows for the regression to be misspecified. Data augmentation, for example, implicitly assumes that the label of the transformed data depends only on the untransformed data, which makes this generalization necessary.*

Detailed assumptions on our data-generating process and model are presented in Sections 2 and 3.

1.2. Summary of Results

The main contributions of our paper are as follows:

(i) *Universality*. Under mild conditions, we prove a set of dependent Gaussian universality results for the training and test risks, which address block dependence (Theorem 2 in Section 3), m-dependence, and specific β-mixing processes (Theorem 3 in Section 4). To the best of our knowledge, this constitutes the first results demonstrating that universality holds in the proportional regime for estimators trained with dependent observations. A key consequence is that if the data is uncorrelated, even if dependent, the asymptotic risk is the same as in the independent setting. Hence previously derived results for logistic regression still hold (see Section 7). To tackle the case where the data is correlated, we propose a novel CGMT result.

(ii) *CGMT*. We introduce a novel extension of the CGMT for Gaussian matrices with a "low-rank" dependence structure (Theorem 5) in Section 5. In particular, this result accommodates dependence across both columns and rows. This significantly broadens the applicability of the CGMT approach, which, until now, required either the rows or the columns to be independent.

(iii) *Data Augmentation*. Using our universality result and the dependent CGMT, we exactly characterize the asymptotic risks of logistic regression under different forms of data augmentation, such as random permutations when the covariates are partially exchangeable and sign flipping when β^* is sparse. We observe that when the structure of the problem is fully known, data augmentation significantly decreases the test risk. However, when it is only partially known, the effect of data augmentation can be negligible. See Section 6.

The remainder of the paper consists of a literature overview in Section 7, an overview of proof techniques in Section 8, and a discussion of future directions in Section 9.

2. Definitions

In this section, we define various quantities that will be used throughout the paper. We first define the empirical risk of an estimator as

$$\hat{R}_n(\beta;\mathbf{X}) \;:=\; \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \omega_i \left(\log\left(1+e^{X_i^\intercal \beta}\right) - y_i X_i^\intercal \beta\right) + \frac{\lambda}{2n}\|\beta\|^2.$$

The performance of our estimator is then evaluated on a new observation X_{new}, which we do not assume to have the same observation as any of the training points. The test risk is hence defined as

$$R_{\text{test}}(\hat{\beta}(\mathbf{X})) \;:=\; \mathbb{E}[\ell_{\text{test}}(X_{\text{new}}^\intercal \hat{\beta}(\mathbf{X}), X_{\text{new}}^\intercal \beta^*) \,\big|\, \hat{\beta}(\mathbf{X})]\;,$$

where the expectation is taken over the mean-zero random vector X_{new} that is independent of the trained estimator $\hat{\beta} = \hat{\beta}(\mathbf{X})$, and where ℓ_{test} is a generic locally Lipschitz function. For our simulations, ℓ_{test} will be the 0-1 loss, for which we also verify our results (see Appendix G). To compare the distribution of training risk on Gaussian and non-Gaussian data, we use the metric given by

$$d_{\mathcal{H}}(X,Y) \;:=\; \sup_{h\in\mathcal{H}} \mathbb{E}\left[h(X)-h(Y)\right],$$

where $\mathcal{H}$ is the set of differentiable functions h with Lipschitz derivative satisfying $\|h\|_\infty, \|h'\|_\infty \leq 1$; see Montanari and Saeed (2022) for why this distance metrizes convergence in distribution.

We shall establish universality with respect to the Gaussian surrogates $G_i \sim \mathcal{N}(0, \mathrm{Var}(X_i))$, where each block $(G_j)_{j\in\mathcal{B}_i}$ is jointly normal. The corresponding dataset $(G_i, y_i(G_i))_{i=1}^n$ satisfies the same assumptions as $(X_i, y_i(X_i))_{i=1}^n$ in Section 3. We also write the Gaussian counterpart of the test risk R_{test} as $R_{\text{test}}^G(\hat\beta(\mathbf{G})) := \mathbb{E}[\ell_{\text{test}}(G_{\text{new}}^\top \hat\beta(\mathbf{G}), G_{\text{new}}^\top \beta^*) \,|\, \hat\beta(\mathbf{G})]$, where $G_{\text{new}} \sim \mathcal{N}(0, \Sigma_{\text{new}})$ for $\Sigma_{\text{new}} := \mathrm{Var}(X_{\text{new}})$ is a substitute for X_{new}.

In our proofs, we will restrict our minimization problem to a particular set of the form

$$\mathcal{S}_p := \left\{\beta \in \mathbb{R}^p : \|\beta\|_2 \leq \mathsf{L}\sqrt{p},\ \|\beta\|_\infty \leq \mathsf{L} p^{\frac{1-r}{2}}\right\} \tag{4}$$

for fixed constants $\mathsf{L} > 0$ and $r \in (0, \frac{1}{8})$. This can be viewed as the set of parameter vectors β which cannot align too strongly with a particular direction to ensure pointwise normality, and is widely used in proving universality results (e.g., Lahiry and Sur (2024); Han and Shen (2023); Montanari and Saeed (2022)). This restriction becomes equivalent to the unconstrained minimization when one proves that $\hat\beta \in \mathcal{S}_p$ with high probability, which can be done on a case-by-case basis. See Appendix D.1 for a more detailed discussion on the set $\mathcal{S}_p$.

3. Universality of the Risks Under Block Dependence

We first state the various assumptions we place on our data generating process and the model. We postpone the discussion of those assumptions to Section 3.1 after the result is stated.

Assumption 1 (Block-dependence) *There exists $k \geq 1$ such that (X_i, y_i) is independent of (X_j, y_j) whenever $j \notin \mathcal{B}_i = \left\{k\lfloor \frac{i-1}{k}\rfloor + 1, \ldots, k\lfloor \frac{i-1}{k}\rfloor + k\right\}$.*

Assumption 2 (Logistic Model) *The labels are generated as $y_i(X_i) = \mathbb{I}(X_i^\top \beta^* - \varepsilon_i > 0)$, where each $\varepsilon_i \sim \mathrm{Logistic}(0,1)$.*

Assumption 3 (Scaling & Sub-Gaussianity) *$\mathbb{E}[X_i] = 0$ and $\mathbb{E}[X_i X_i^\top] = \Sigma_i$. Moreover, each X_i is sub-Gaussian, and there exists $\mathsf{K}_X > 0$ such that $\sup_{1\leq i\leq n} \|X_i\|_{\psi_2} \leq \mathsf{K}_X/\sqrt{n}$, where $\|\bullet\|_{\psi_2}$ denotes the sub-Gaussian norm.*

Assumption 4 (Signal Size) *$\beta^* \in \mathcal{S}_p$ as in (4), and there exists $\kappa \in (0,\infty)$ such that $\frac{p}{n} = \frac{p(n)}{n} \to \kappa$.*

Assumption 5 (Gaussian Approximation) *For the hypersphere $\mathcal{S}^{k-1} := \{\boldsymbol{x} \in \mathbb{R}^k : \|\boldsymbol{x}\|_2 = 1\}$,*

$$\sup_{f\in\mathcal{F}} \sup_{\beta_1,\ldots,\beta_k\in\mathcal{S}_p} \sup_{\theta\in\mathcal{S}^{k-1},\, i\leq n-k} \left|\mathbb{E}\Big[f\Big(\textstyle\sum_{r=1}^k \theta_r X_{i+r}^\top \beta_r\Big) - f\Big(\textstyle\sum_{r=1}^k \theta_r G_{i+r}^\top \beta_r\Big)\Big]\right| \;\to\; 0\,,$$

where $\mathcal{F} := \{f : \mathbb{R} \to \mathbb{R} \mid f \in C_1, \|f\|_\infty < \infty, \|\partial f\|_\infty \leq 1\}$.

Assumptions 1-5 are used to establish universality of the training risk. For universality of the test risk, we require two additional assumptions: one on the distribution of X_{new}, and one on the geometry of the training risk. Below, we denote $\tilde{\mathcal{F}} := \{f : \mathbb{R}^2 \to \mathbb{R} \mid f \in C_1, \|f\|_\infty < \infty, \|\partial f\|_\infty \leq 1\}$.

Assumption 6 (Gaussian Approximation of $X_{\mathbf{new}}$) *We have*

$$\sup_{f\in\tilde{\mathcal{F}}}\ \sup_{\beta\in\mathcal{S}_p}\ \left|\mathbb{E}[f(X_{\text{new}}^T\beta, X_{\text{new}}^T\beta^*) - f(G_{\text{new}}^T\beta, G_{\text{new}}^T\beta^*)]\right| \ \to\ 0\ .$$

Assumption 7 *There exist constants $\bar{\chi}, \chi_* > 0$ such that for every fixed $\epsilon > 0$,*

$$\mathbb{P}\left(\min_{\beta\in\mathcal{S}_p\,,\,|(\beta^\top\Sigma_{new}\beta)^{1/2}-\bar{\chi}|>\epsilon}\hat{R}_n(\beta;\mathbf{G})\ >\ \min_{\beta\in\mathcal{S}_p}\hat{R}_n(\beta;\mathbf{G})\right)\ \to\ 1 \qquad \text{and} \qquad \beta^{*\top}\Sigma_{new}\beta^*\ \to\ \chi_*^2\ .$$

Under these assumptions, the following theorem holds:

Theorem 2 (Block Dependent Universality) *Let $(X_i, y_i(X_i))_{i=1}^n$ and $(G_i, y_i(G_i))_{i=1}^n$ be generated under Assumptions 1-5, where each $G_i \sim \mathcal{N}(0, \mathrm{Var}(X_i))$. Then*

$$d_{\mathcal{H}}(\ \min_\beta \hat{R}_n(\beta;\mathbf{X}), \min_\beta \hat{R}_n(\beta;\mathbf{G})) \to 0. \tag{5}$$

Moreover, if Assumptions 6 and 7 also hold, then

$$|\,R_{\text{test}}(\hat{\beta}(\mathbf{X})) - R_{\text{test}}^G(\hat{\beta}(\mathbf{G}))\,|\ \xrightarrow{\mathbb{P}}\ 0\ . \tag{6}$$

The proof of Theorem 2 is deferred to Appendix D and G, with a proof sketch given in Section 8. This result allows us to better understand the properties of the risk—notably, we observe that it only depends on the distribution of the data through its first two moments. Consequently, the dependence among the observations influences the risk only via the covariance, rather than through a more intricate relationship. In other words, even if the data exhibits dependence, as long as it is uncorrelated, the asymptotic behavior of the risk is the same as the independent case, and allows it to be analyzed using the existing extensive literature. In scenarios where the data is not uncorrelated, the risk of $\hat{\beta}(\mathbf{X})$ still simplifies to the risk of $\hat{\beta}(\mathbf{G})$, and such cases can be studied using our novel dependent CGMT approach (see Section 5), as long as a "low rank" dependence assumption holds.

3.1. Discussion of Assumptions

Assumptions 2–4 are standard in high-dimensional settings. Assumptions 5–7 are mirror conditions required to establish universality of the risk in the independent case (e.g. Montanari and Saeed (2022); Han and Shen (2023)). In particular, our Assumption 5 is closely related to Assumption 5 of Montanari and Saeed (2022). However, ours is slightly stronger, as it requires the joint convergence of $(X_{i_1}^T\beta_1, \ldots, X_{i_k}^T\beta_k)$ to a Gaussian limit for all $\beta_1, \ldots, \beta_k \in \mathcal{S}_p$. This is a direct consequence of Assumption 1, which relaxes the independence assumption to block dependence. It can hence be seen as a multivariate version of the pointwise normality assumption in Montanari and Saeed (2022).

To establish the Gaussian universality of the testing risk, in addition to the training risk, it is necessary to introduce further assumptions, specifically Assumptions 6 and 7. Assumption 6 closely resembles Assumption 5, but applies to X_{new} rather than our original data. Note that we did not require X_{new} to share the same distribution as any of the (X_i). Assumption 7 is a stronger condition: informally, it states that in the Gaussian case, the optimizer should be concentrated on a small subset of $\mathcal{S}_p$. However, since this assumption pertains to the Gaussian data rather than $\mathbf{X}$, it can be proven via our dependent CGMT framework, provided that Assumption 10 is satisfied. In the independent setting this is notably established in Salehi et al. (2019, Eq. 92), Dhifallah and Lu (2021, Eq. 74) and Thrampoulidis (2016, Eq. B.11). Montanari and Saeed (2022) does not impose such a condition, but instead studied a modified notion of the test risk. More formally Montanari and Saeed (2022) proved the universality of $\min_{\beta\in\tilde{\mathcal{S}}(\mathbf{X})} R_{\text{test}}(\beta)$ where $\tilde{\mathcal{S}}(\mathbf{X}) \subset \mathcal{S}_p$ is a subset defined using the empirical risk (see Theorem 2 of Montanari and Saeed (2022)).

4. Extension to m-dependent and Specific β-Mixing Processes

In this section, we show that the universality result of Theorem 2 also extends to data with m-dependence and specific β-mixing processes. In the definition of mixing, we shall temporarily make the dimension dependence explicit in the data $(X_i, y_i) \equiv (X_i^{(p)}, y_i^{(p)})$ and $\mathbf{X} \equiv \mathbf{X}^{(p)}$. Following Bradley (2005), for every $p \in \mathbb{N}$, we define the β-mixing coefficient of $(X_i^{(p)}, y_i^{(p)})_{i\in\mathbb{N}}$ as

$$\beta_{\text{mix}}(N) \;:=\; \sup_{p\in\mathbb{N}}\sup_{t\in\mathbb{N}} \sup_{\mathcal{A}\in\mathcal{P}^{(p)}_{\le t},\,\mathcal{B}\in\mathcal{P}^{(p)}_{\ge t+N}} \frac{1}{2}\sum\nolimits_{A\in\mathcal{A}}\sum\nolimits_{B\in\mathcal{B}}\big|\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)\big|\,,$$

where $\mathcal{P}^{(p)}_{\le t}$ is the set of all finite partitions of $\sigma((X_i^{(p)}, y_i^{(p)})\,|\, i \le t)$ and $\mathcal{P}^{(p)}_{\ge t+N}$ is the set of all finite partitions of $\sigma((X_i^{(p)}, y_i^{(p)})\,|\, i \ge t+N)$. We use this definition to state our β-mixing requirement for the triangular array $(X_i^{(p)}, y_i^{(p)})_{i,p\in\mathbb{N}}$ below in Assumption 8 (ii).

Assumption 8 (m-dependence or specific β-mixing processes) *One of the following holds:*

(i) (X_i, y_i) *is independent of* (X_j, y_j) *whenever* $|i-j| > m$*;*

(ii) There exists some $r \in (0,1)$ *such that* $\sum_{l=1}^{\infty}\beta_{\text{mix}}(l)^r < \infty$*. Moreover, there exists a process* $(Z_i^{(p)})_{i,p\in\mathbb{N}}$ *of centered independent random vectors with bounded sub-Gaussian norms such that, for every* $i, p \in \mathbb{N}$*, we can express*

$$X_i^{(p)} \;=\; \sum\nolimits_{j\in\mathbb{N}} c_{i,j}^{(p)}\, Z_j^{(p)}$$

for some constants $(c_{i,j}^{(p)})$*. We also assume that there exists an universal constant* $\underline{c} > 0$ *such that* $\sqrt{p}\lambda_{\min}(\text{Var}(Z_j^{(p)})) \ge \underline{c}$ *for all* $j, p \in \mathbb{N}$*.*

Assumption 8 substitutes the block-dependence Assumption 1. Assumption 8(ii) is restrictive, but already covers important β-mixing processes of practical interests: For example, any autoregressive model (Tsay, 2005; Gelfand et al., 2010) can be expressed in the form of Assumption 8(ii).

Assumption 9 (Strong Gaussian Approximation) *For all* $d \in \mathbb{N}$*, we have*

$$\sup_{f\in\mathcal{F}}\ \sup_{\beta_1,\ldots,\beta_d\in\mathcal{S}_p}\ \sup_{\theta\in\mathcal{S}^{d-1},\,i\le n-d}\ \Big|\mathbb{E}\Big[f\Big(\sum\nolimits_{r=1}^{d}\theta_r X_{i+r}^{\intercal}\beta_r\Big) - f\Big(\sum\nolimits_{r=1}^{d}\theta_r G_{i+r}^{\intercal}\beta_r\Big)\Big]\Big| \;\to\; 0\,,$$

where $\mathcal{F} := \{f : \mathbb{R}\to\mathbb{R} \mid f \in C_1, \|f\|_\infty < \infty, \|\partial f\|_\infty \le 1\}$*.*

Assumption 9 is a stronger form of Assumption 5: Instead of requiring the Gaussian approximation to hold for k-many vectors, where k is the dependency block size, we now require this to hold for every fixed $d \in \mathbb{N}$. Note however that we do not require d to grow with n; it suffices that for every fixed d, the convergence holds as $n \to \infty$. By replacing Assumptions 1 & 5 by Assumptions 8 & 9, we may state our extended universality result for m-dependence and specific mixing processes.

Theorem 3 (Universality under m-dependence or mixing) *Let* $(X_i, y_i(X_i))_{i=1}^n$ *and* $(G_i, y_i(G_i))_{i=1}^n$ *be generated under Assumptions 2-4, where each* $G_i \sim \mathcal{N}(0, \text{Var}(X_i))$*. Assume in addition that Assumptions 8 and 9 hold. Then*

$$d_{\mathcal{H}}(\min\nolimits_\beta \hat{R}_n(\beta;\mathbf{X}), \min\nolimits_\beta \hat{R}_n(\beta;\mathbf{G})) \to 0. \tag{7}$$

Moreover, if Assumptions 6 and 7 also hold, then

$$|R_{\text{test}}(\hat{\beta}(\mathbf{X})) - R^{G}_{\text{test}}(\hat{\beta}(\mathbf{G}))| \overset{\mathbb{P}}{\to} 0\ . \qquad (8)$$

The proof of Theorem 3 is an adaptation of the block-dependent case of Theorem 2, and is included in Appendices E and F. To give an overview, the first step of the proof is to apply the classical technique of representing the data as an alternating sequence of big blocks and small blocks of random vectors, where the small blocks are then ignored (Bernstein, 1927; Ibragimov, 1975; Davidson, 1992). In the m-dependent case, the big blocks become independent provided that each small block is of size at least m, and the block-dependent result of Theorem 2 applies directly. In the mixing case, the big blocks are only approximately independent and, hence, Theorem 2 does not directly apply. To be able to use the results we derived in the independent case, we use the embedding result of Yu (1994). This result allows one to compare the expectation of functions of β-mixing block with that of exactly independent random variables. However, a difficulty is that in our setting, the sizes of the blocks are not allowed to depend on n, and the number of blocks that can be approximated by independent ones cannot depend on n. Hence, we cannot apply the embedding result of Yu (1994) directly to the risk. We instead use Yu (1994) to bound how much the risk changes along the path of interpolation between $\mathbf{G}$ and $\mathbf{X}$. Finally, note that a key ingredient of the proof is being able to control the largest eigenvalue of $\mathbf{X}^T\mathbf{X}$. The proof relies on being able to prove a Bernstein inequality for $\sum_{i=1}^{n}(X_i^T\beta)^2$ for all $\beta \in \mathcal{S}_p$. Under the mixing setting, Assumption 8(ii) allows us to do so by rewriting $\sum_{i=1}^{n}(X_i^T\beta)^2$ as a quadratic form over the independent process (Z_j). In general, this assumption — that (X_i) can be rewritten as an infinite sum of independent processes — can be weakened to processes for which one can control the moments of $\lambda_{\max}(\mathbf{X}^T\mathbf{X})$.

5. Dependent CGMT

Under general conditions, Theorems 2 and 3 allows us to study the risk of $\hat{\beta}(\mathbf{X})$ via that of $\hat{\beta}(\mathbf{G})$. When X_i's are isotropic and uncorrelated, $\mathbf{G} = (G_i)_{i\leq n}$ can be viewed as an $\mathbb{R}^{p\times n}$ matrix with i.i.d. standard Gaussian entries. In this case, the risk of $\hat{\beta}(\mathbf{G})$ can be studied via the CGMT method. Broadly speaking, the classical CGMT method first relates $\min_\beta \hat{R}_n(\beta;\mathbf{G})$ to the optimization

$$\Psi_{\mathcal{S}_w,\mathcal{S}_u} := \min_{w\in\mathcal{S}_w} \max_{u\in\mathcal{S}_u} L_\Psi(w,u) \qquad \text{with} \qquad L_\Psi(w,u) := w^\intercal \mathbf{H} u + f(w,u)\ , \qquad (9)$$

where $\mathcal{S}_w \subset \mathbb{R}^p$ and $\mathcal{S}_u \subset \mathbb{R}^n$ are compact and convex, $f : \mathcal{S}_w \times \mathcal{S}_u \to \mathbb{R}$ is a convex-concave function, and $\mathbf{H}$ is typically a suitably projected version of $\mathbf{G}$. The main result of CGMT is that $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$ is equivalent to a simpler optimization involving only two Gaussian vectors (see Appendix B.2).

In the dependent setting, however, $\mathbf{H}$ can exhibit both correlated columns and rows, and the standard CGMT framework is generally not applicable. To address this, we develop a more general CGMT framework that accommodates a "low-rank assumption" on the dependence structure of $\mathbf{H}$.

Assumption 10 (Low-rank Dependence) *Let* $\mathbf{H}$ *be an* $\mathbb{R}^{p\times n}$ *Gaussian matrix. There exist* $M \in \mathbb{N}$ *and symmetric positive semi-definite matrices* $(\Sigma^{(l)}, \tilde{\Sigma}^{(l)})_{l\leq M}$*, with* $\Sigma^{(l)} \in \mathbb{R}^{p\times p}$ *and* $\tilde{\Sigma}^{(l)} \in \mathbb{R}^{n\times n}$*, s.t.*

$$\operatorname{Cov}[\mathbf{H}_{ji}, \mathbf{H}_{j'i'}] = \sum_{l=1}^{M} \Sigma^{(l)}_{jj'} \tilde{\Sigma}^{(l)}_{ii'} \qquad \textit{for all } i,i' \leq n \textit{ and } j,j' \leq p\ .$$

Remark 4 *For* $M = 1$*, Assumption 10 is equivalent to the assumption that* $\mathbf{H} = \Sigma^{(1)}\mathbf{H}'\tilde{\Sigma}^{(1)}$ *for some Gaussian matrix* $\mathbf{H}'$ *with i.i.d. standard normal entries. For* $M > 1$*, this says that the dependence* *is captured by some sum of covariance matrices that "factorise". In Appendix B.4, we discuss how this is a natural assumption for the dependence structure that arises in data augmentation.*

Denote $\|v\|_{\Sigma'} = \sqrt{v^\intercal \Sigma' v}$. Under Assumption 10, we shall compare $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$ to the risk

$$\psi_{\mathcal{S}_w,\mathcal{S}_u} \;:=\; \min_{w\in\mathcal{S}_w}\; \max_{u\in\mathcal{S}_u}\; L_\psi(w,u)\;,$$
$$\text{where}\quad L_\psi(w,u) \;:=\; \textstyle\sum_{l=1}^M \Big\{\|w\|_{\Sigma^{(l)}} \mathbf{h}_l^\intercal (\tilde{\Sigma}^{(l)})^{1/2} u + w^\intercal (\Sigma^{(l)})^{1/2} \mathbf{g}_l \|u\|_{\tilde{\Sigma}^{(l)}}\Big\} + f(w,u)\;. \qquad (10)$$

$(\mathbf{h}_l, \mathbf{g}_l)_{l\leq M}$ are independent standard Gaussians respectively in $\mathbb{R}^n$ and $\mathbb{R}^p$. Our next result formalizes the equivalence of $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$ and $\psi_{\mathcal{S}_w,\mathcal{S}_u}$, and additionally controls $\hat{w}_\Psi \in \mathcal{S}_p$, the minimizer of $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$.

Theorem 5 (Dependent CGMT) *Suppose $\mathcal{S}_w$ and $\mathcal{S}_u$ are compact and convex, and f is continuous and convex-concave on $\mathcal{S}_w \times \mathcal{S}_u$. Under Assumption 10, the following statements hold:*

(i) For all $c \in \mathbb{R}$,

$$\mathbb{P}(\Psi_{\mathcal{S}_w,\mathcal{S}_u} \leq c) \;\leq 2^M\, \mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \leq c) \qquad \textit{and} \qquad \mathbb{P}(\Psi_{\mathcal{S}_w,\mathcal{S}_u} \geq c) \;\leq\; 2^M\, \mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \geq c)\;;$$

(ii) Let $\mathcal{A}_p$ be an arbitrary open subset of $\mathcal{S}_w$ and $\mathcal{A}_p^c := \mathcal{S}_w \setminus \mathcal{A}_p$. If there exist constants $\bar{\phi}_{\mathcal{S}_w}$, $\bar{\phi}_{\mathcal{A}_p^c}$ and $\eta, \epsilon > 0$ such that $\bar{\psi}_{\mathcal{A}_p^c} \geq \bar{\psi}_{\mathcal{S}_w} + 3\eta$, $\mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \leq \bar{\psi}_{\mathcal{S}_w} + \eta) \;\geq\; 1-\epsilon$ and $\mathbb{P}(\psi_{\mathcal{A}_p^c,\mathcal{S}_u} \geq \bar{\psi}_{\mathcal{A}_p^c} - \eta) \;\geq\; 1-\epsilon$, then

$$\mathbb{P}(\hat{w}_\Psi \;\in\; \mathcal{A}_p) \;\geq\; 1-4\epsilon\;.$$

Remark 6 *Convexity is not required for the first bound of (i); see Theorem 13 for the full theorem.*

Notably, Theorem 5 implies an asymptotic concentration result for the minimizer $\hat{w}_\Psi$ in $\mathcal{A}_p$:

Corollary 7 (Asymptotic CGMT) *Let $\mathcal{A}_p$ be an arbitrary open subset of $\mathcal{S}_w$ and $\mathcal{A}_p^c := \mathcal{S}_w \setminus \mathcal{A}_p$. If there exists constants $\bar{\psi} < \bar{\psi}^c$ such that $\psi_{\mathcal{S}_w,\mathcal{S}_u} \xrightarrow{\mathbb{P}} \bar{\psi}$ and $\psi_{\mathcal{A}_p^c,\mathcal{S}_u} \xrightarrow{\mathbb{P}} \bar{\psi}^c$, then*

$$\mathbb{P}(\hat{w}_\Psi \in \mathcal{A}_p) \;\rightarrow\; 1\;.$$

Theorem 5 and Corollary 7 have several important implications:

Simplifying the analysis of $\hat{R}_n(\beta;\mathbf{G})$. Theorem 5 reduces the analysis of $\hat{R}_n(\beta;\mathbf{G})$, which involves a high-dimensional and correlated Gaussian matrix, to a loss involving only Gaussian vectors. This substantially simplifies the analysis of the asymptotic risk, as one can avoid invoking random matrix theory. Indeed in the isotropic case, the conversion of $\psi_{\mathcal{S}_w,\mathcal{S}_u}$ into a deterministic, low-dimensional problem has been performed in many models through algebraic calculations and the min-max theorem (Thrampoulidis, 2016; Salehi et al., 2019; Dhifallah and Lu, 2021). In our case, the terms in $\psi_{\mathcal{S}_w,\mathcal{S}_u}$ depend on the $2M$ covariance matrices, and the complexity of these calculations grows with M. We present the calculations for special cases of data augmentation in Section 6 and Appendix B.

Pipeline of analysis for dependent data. Together with our dependent universality result (Theorems 2 and 3), Theorem 5 extends the pipeline of analysis discussed in Section 1 to dependent data in logistic regression. Since Theorem 5 is model-independent, we also expect it to be valuable to other setups, provided that an analogous dependent universality result is established.

Universality of test risk. Our CGMT also helps with verifying Assumption 7, required for the universality of test risk in Theorems 2 and 3. To see this, let us identify $\mathcal{A}_p$ in Theorem 5(*ii*) as the set $\{\beta \in \mathcal{S}_p \mid |(\beta^\top \Sigma_{\text{new}} \beta)^{1/2} - \bar{\chi}| \leq \epsilon_n\}$. Under this notation, Assumption 7 is a comparison between the training risks of two optimizations on $\mathcal{S}_w \setminus \mathcal{A}_p$ and $\mathcal{S}_w$ respectively. Theorem 5 allows us to perform this comparison on the simpler auxiliary optimizations instead, which additionally allows for computing the value of $\bar{\chi}$; see Appendix B.1.

Remark 8 (Comparison to existing CGMT results) *For comparison, Theorem 5 recovers the standard CGMT with $\Sigma^{(1)} = I_p$, $\tilde{\Sigma}^{(1)} = I_n$ and $M = 1$. It also recovers the multivariate CGMT of Dhifallah and Lu (2021) by setting $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$ as block diagonal matrices with M equal-sized subblocks, such that the l-th subblock is identity and the other blocks are zero. Akhtiamov et al. (2024a) generalizes the block diagonal setup to allow non-identity subblocks, which is a special case of our Assumption 10, but they also allow for transforming w and u, which we do not address here.*

6. Applications to Data Augmentation (DA)

As an example application of Theorems 2 and 5, we analyze the effect of data augmentation, which introduces a simple yet ubiquitous form of block dependence in machine learning. To see how the dependence arises, let $(Z_i)_{i \leq m}$ be i.i.d. mean-zero random vectors, which correspond to our original data, and let $\phi_1, \ldots, \phi_{mk}$ be $n = mk$ i.i.d. $\mathbb{R}^p \to \mathbb{R}^p$ transformations, which are the augmentations. Note that the coordinates of Z_i may be locally dependent. DA synthesizes an artificial dataset $(X_i, y_i)_{i \leq n}$ by setting $X_i \coloneqq \phi_i(Z_{\lceil i/k \rceil})$, i.e. each observation is augmented k times, and setting $y_i = y_i(Z_{\lceil i/k \rceil})$ to retain the label of the original observation. The estimator $\hat{\beta}$ is then fitted on the augmented dataset through the minimization

$$\min\nolimits_{\beta \in \mathcal{S}_p} \frac{1}{n} \sum\nolimits_{i=1}^{n} \left(\log\left(1 + e^{X_i^\top \beta}\right) - y_i(Z_{\lceil i/k \rceil}) \times X_i^\top \beta \right) + \frac{\lambda}{2n} \|\beta\|_2^2 \,. \tag{11}$$

See Remark 1 for how this relates to the model (3). For simplicity, we assume that the test data X_{new} is identically distributed as the unaugmented Z_1. The practical heuristic behind DA is that, if ϕ_i's are chosen to reflect certain structures of the problem well, DA may improve the test risk of $\hat{\beta}$ despite the dependence introduced. While the benefits of DA are empirically observed across a large body of ML literature, limited theoretical attempts have provided an exact theoretical quantification, especially in the case of a classification task; see Section 7. Here, we analyze several DA schemes:

Random permutations under a group structure. Suppose Z_1 can be broken down into N groups of coordinates, each of size p_t, with $p_1 + \ldots + p_N = p$. Namely, $Z_1 = ((Z_1^{(1)})^\top, \ldots, (Z_1^{(N)})^\top)^\top$ such that $\{Z_1^{(t)}\}_{t \leq N}$ are independent vectors and each $Z_i^{(t)}$ is $\mathbb{R}^{p_t}$-valued with i.i.d. coordinates. The i.i.d. structure within each group motivates one to augment the data by permuting coordinates within each group. As the full permutation group is exponentially large in p, a practical question is how much permutation should one perform. This concerns both the number of random permutations k as well as the proportion of coordinates to permute. For simplicity, we fix $r_{\text{perm}} \in [0, 1]$, a proportionality parameter that may be chosen by practitioners, so that within each t-th group, we only consider permuting the top $\lceil r_{\text{perm}} p_t \rceil$ coordinates. Each augmentation ϕ_i is a uniformly random permutation that permutes the top $\lceil r_{\text{perm}} p_t \rceil$ coordinates within each t-th group.

Random sign flipping under a sparsity structure. Consider a sparsity structure in β^*: A proportion $\rho^* \in [0, 1]$ of the p entries of β^* are non-zero, whereas the remaining $(1 - \rho^*)p$ entries are

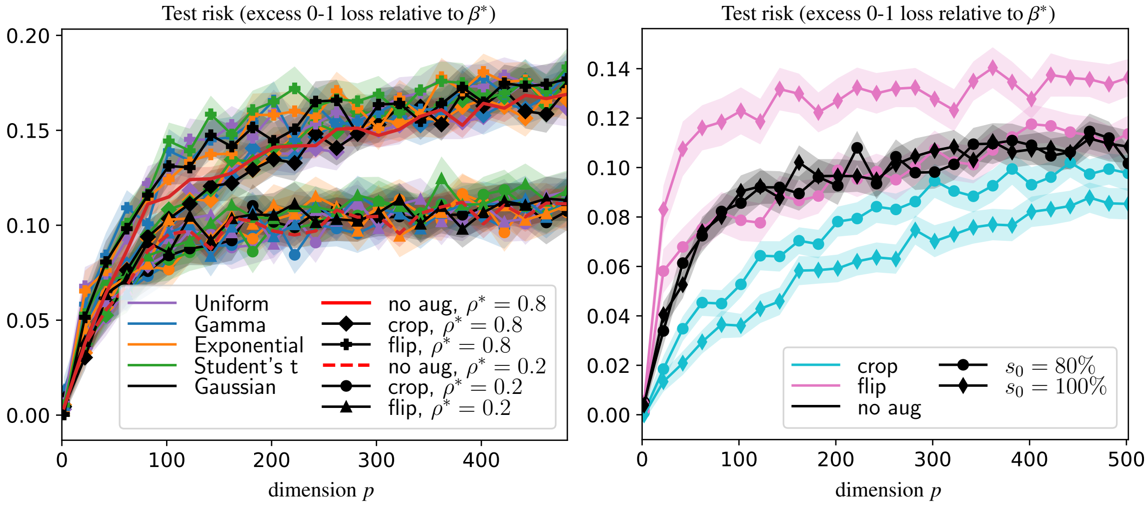

Figure 2: Test risks under random cropping and sign flipping. *Left*. Same setup as Section 6. *Right*. Signal ratio $\rho^* = 0.2$ and the bottom $\lceil s_0(1-\rho^*)p \rceil$ coordinates are known to be zero.

zero. The positions of the non-zero coordinates are unknown in general, and ρ^* may be known or unknown. This motivates the use of random sign flipping to shrink the estimate $\hat{\beta}$ at locations where the entries of β^* may be zero. We fix some $\lceil r_{\text{flip}} p \rceil$ entries of p, where $r_{\text{flip}} \in [0,1]$ is a parameter chosen by users. Each ϕ_i is a random diagonal matrix, generated by drawing the fixed $\lceil r_{\text{flip}} p \rceil$ entries of its diagonal as i.i.d. Rademacher variables, and setting the remaining entries of the diagonal to 1.

Random cropping under a sparsity structure. Suppose β^* has the same sparsity structure as above. Another way to encode our guess of the zero entries is by randomly removing coordinates in the data. Here, each ϕ_i is a random diagonal matrix generated by randomly setting $\lceil r_{\text{crop}} p \rceil$ entries of its diagonal to zero and leaving the remaining entries as 1.

As y_i's depend on the unaugmented data instead of the augmented, we apply a generalization of Theorem 2 (Appendix D.2) to show that universality holds under the same assumptions. The key condition to verify is Assumption 5. In Appendix J, we verify this for sign flipping, cropping and noise injection. For permutations, we verify this under general conditions on the group sizes p_t's for both a fixed and small number of groups N and a growing number of groups $N = N(n) \to \infty$. Meanwhile, our CGMT result (Theorem 5) applies to sign flipping and cropping above with $\mathrm{Var}[Z_1] = \frac{1}{p} I_p$, as well as permutations (Appendix M.2). Under simplifying conditions that hold for permutations and sign flipping, we also derive a set of 10 deterministic and scalar equations (EQs) in Appendix B.1, which explicitly characterize the test risk of logistic regressors. More general augmentations can be accommodated but at the expense of a more complicated system of equations. While (EQs) is complicated to state, we verify that in the isotropic case with no augmentation, it recovers exactly the characterizing equation by Salehi et al. (2019).

These results enable us to understand the effects of DA through two possible approaches: To simulate a logistic regression on Gaussians, i.e. a high-dimensional convex optimization with simple distributions, or to solve the nonlinear equations (EQs), i.e. a low-dimensional but highly non-convex optimization. We use the former approach with $m = 200$ synthetic data and present results in Fig. 1 ($p = 500$) and 2 ($k = 30$); see Appendix C for full simulation details[†]. A few observations:

[†]. Code is available at github.com/KevHH/Dependent_Universality.

Full permutations alleviate overfitting under a group structure. Fig. 1 considers a high dimensional regime ($p/m = 2.5$), where logistic regression with no augmentations is expected to overfit. This typically manifests through a low training risk but a high test risk. Fig. 1 shows that a full permutation ($r_{\text{perm}} = 1.0$) of the i.i.d. coordinates guards against this overfitting: The test risk improves substantially and as more and more augmentations are used.

Full knowledge of the problem structure can be crucial. A surprising observation from Fig. 1 is that using only a slightly smaller subgroup of permutations ($r_{\text{perm}} = 0.8$) results in test risks that are within error margins from that of no augmentations. This suggests that, at least within our model, exploiting the full set of permutation invariance is critical for obtaining noticeable improvements. On the other hand, the sparsity setup for cropping and sign flipping does not allow the knowledge of the full structure by design, as it would otherwise imply that we know exactly which coordinates of β to exclude from the regression. In the left plot of Fig. 2, perhaps surprisingly, we see that sign flipping and cropping both yield indistinguishable test risks from that under no augmentation. For comparison, we perform Gaussian simulation in an artificial setup in the right plot of Fig. 2, where some portion of the null entries of β^* are known and on which cropping and sign flipping are always performed. The remaining amount of cropping and sign flipping are applied to the rest of the coordinates. There, cropping ensures that the known null coordinates never enter the regression and outperforms no augmentation and sign flipping. In summary, these observations send a cautionary message: the benefits of data augmentation may be concretely visible only when the full problem structure is known, which is too stringent for many practical setups.

7. Related Literature

Universality. Universality has been extensively studied in the probability, statistics, and ML literature. In statistics, the risk of a wide range of penalized linear models and the behavior of the approximate message passing (AMP) algorithm have been demonstrated to be universal (Korada and Montanari, 2011; Montanari and Nguyen, 2017; Abbasi et al., 2019; Oymak and Tropp, 2018; Han and Shen, 2023; Dudeja et al., 2023; Wang et al., 2024; Chen and Lam, 2021). Beyond linear regression, it has been proven that generalized linear models, perceptron models, max-margin classifiers, random feature models, and others obtained via empirical risk minimization exhibit universal behavior (Montanari and Saeed, 2022; Montanari et al., 2023; Dandi et al., 2024; Gerace et al., 2024; Korada and Montanari, 2011; Han and Shen, 2023; Hu and Lu, 2022). Those works either assume that the covariates are independent or that the observations projected on a wide range of directions are asymptotically normal (e.g., Montanari and Saeed (2022)). Lahiry and Sur (2024) further proved it for regularized linear regression if the covariates within each vector are block-dependent. However, they still assumed the rows of the design matrix were independent. Huang et al. (2022) moved beyond this condition and showed that under certain stability conditions, machine learning estimators trained with data augmentation satisfy Gaussian universality. Those conditions are, however, hard to verify and this paper does not cover overparameterized logistic regression.

CGMT & Exact Asymptotics. The exact asymptotic risk of many high-dimensional models has been extensively studied using a variety of techniques. These include AMP (Donoho et al., 2009), the cavity method (Opper et al., 2001), the Gaussian Min-Max Theorem (Gordon, 1985), and the CGMT (Thrampoulidis et al., 2014). Since its introduction, the CGMT has been successfully applied to analyze the risk of numerous high-dimensional models (e.g. Stojnic (2013a,b); Thrampoulidis et al. (2015, 2018); Akhtiamov et al. (2024b); Aolaritei et al. (2022); Javanmard and

Soltanolkotabi (2022); Mignacco et al. (2020)). Further developments have extended the method to settings with independent but non-identically distributed rows (Akhtiamov et al., 2024a; Dhifallah and Lu, 2021). We, for the first time, extend CGMT to dependent rows and columns.

Data Augmentation. Data augmentation is a widely utilized practice in machine learning, particularly in deep learning (e.g. Taqi et al. (2018); Shorten and Khoshgoftaar (2019); Shorten et al. (2021); Volkova (2024)). Given its critical role, a number of studies have investigated its theoretical properties (e.g. Hanin and Sun (2020); Huang et al. (2022); Chen et al. (2020); Lin et al. (2024)). The first work that applies CGMT to the study of data augmentation is Dhifallah and Lu (2021), which examines the impact of noise injection on logistic regression, demonstrating that it serves as an implicit regularization. However, their results and analysis are limited to noise injection, a data augmentation strategy that preserves the independence of the covariates, which simplifies their study. In this paper, we develop a novel universality and CGMT result that allows us to significantly broaden the scope of data augmentations we can study.

Logistic Regression. Recently, substantial progress has been made in understanding the exact asymptotics of high-dimensional logistic regression in the proportional regime (Sur and Candès, 2019; Deng et al., 2022; Kini et al., 2021). Salehi et al. (2019) successfully adapted the CGMT framework to the logistic regression setting, enabling the analysis of regularized logistic regression. The issue of dependence in logistic regression, motivated by applications to biology and sociology, has also been well-studied (Bonney, 1987; Prentice, 1988; Reboussin et al., 2008; Zorn, 2001). The dependence was notably modeled through mixed effects or latent variable models. In high dimensions, recent work has taken inspiration from the Ising model to propose a model in which y_i depends not only on X_i but also other labels $(y_j)_{j\neq i}$, exhibiting network dependence (Mukherjee et al., 2021). Beyond the locally dependent setting, a recent wave of papers has studied properties of estimators trained on dependent data (Nagaraj et al., 2020; Zou et al., 2009). Beyond logistic regression, a number of papers have studied the asymptotics of Lasso estimators trained on time series data and in VAR models (e.g Wong et al. (2020); Basu and Michailidis (2015); Wong et al. (2016); Nicholson et al. (2020); Guo et al. (2016)).

8. Proof Overview

Universality. The proof for the training risk builds upon a variant of the Lindeberg method (see e.g. Chatterjee (2005)) introduced by Montanari and Saeed (2022). One crucial difference, however, is that we need to account for the dependence between the observations. In Montanari and Saeed (2022), the independence of the data allows one to reduce the proof to showing that the mean of a particular function of $X_i^\top\beta$ approaches zero for all $\beta \in \mathcal{S}_p$. This is done by exploiting the asymptotic normality of $X_i^\top\beta$. However, in the presence of dependence, this reduction is no longer valid. Instead, we must control the mean of a function of $(X_{i_1}^\top\beta_1, \dots, X_{i_k}^\top\beta_k)$ for all $\beta_1, \dots, \beta_k \in \mathcal{S}_p$. This requires establishing its joint asymptotic normality and a more careful analysis.

The proof for the test risk builds on the observation that under Assumption 6, the test risk depends asymptotically only on $\hat{\beta}(\mathbf{X})^T\Sigma_{\text{new}}\hat{\beta}(\mathbf{X})$. We further exploit universality of the training risk and Assumption 7 to obtain that $\hat{\beta}(\mathbf{X})^T\Sigma_{\text{new}}\hat{\beta}(\mathbf{X})$ converges in probability to a deterministic constant $\bar{\chi}^2$, which is the same limit as in the Gaussian case. The desired result then directly follows.

CGMT for Data Augmentation (DA). In DA, the covariance of a set of augmented data $\{\phi_1(Z_1), \dots, \phi_k(Z_1)\}$ is completely described by the variance of the individual data points $\mathrm{Var}[\phi_1(Z_1)]$ and the covariance between two differently transformed data $\mathrm{Cov}[\phi_1(Z_1), \phi_2(Z_1)]$. As a result, this satisfies the low-rank dependence assumption of our CGMT (Assumption 10) with $M = 2$. However, the actual application of the CGMT is more subtle since the logistic regression (3) is not a priori in the form of the primary optimization (9). Similar to Thrampoulidis (2016); Salehi et al. (2019); Dhifallah and Lu (2021), we first move the data X_i outside the logarithm in (3) via Lagrange multipliers. This yields a formulation similar to (9) involving a high-dimensional $\mathbb{R}^{p\times n}$ matrix $\mathbf{X} = (X_1, \ldots, X_n)$. Due to the presence of the labels y_i and their nonlinear dependence on the data, CGMT can only be applied to a suitably projected version of $\mathbf{X}$, say $P\mathbf{X}$, that is uncorrelated with both the labels y_i and the remainder $(I_p - P)\mathbf{X}$. In the isotropic and independent case, PX_i and $(I_p - P)X_j$ are uncorrelated for any projection matrix P, whereas y_i depends on X_i only through $X_i^\intercal \beta^*$, so one may choose P to project onto the subspace orthogonal to β^*. In the dependent case, PX_i and $(I_p - P)X_j$ may still be correlated, as $\mathrm{Cov}[X_i, X_j]$ does not necessarily commute with P. In DA, this is further complicated by the fact that y_i depends on the unaugmented data Z_i instead of X_i. Choosing a suitable projection P is thus highly non-trivial, and we develop such a P for DA in Appendix M.

9. Conclusion and Future Work

We have shown that, for high-dimensional logistic regression, the pipeline of analysis of Gaussian universality and CGMT extends readily to dependent data, and the asymptotic risks are again completely characterized by the mean and the variance of the data. This has many useful implications, from allowing us to perform Gaussian simulations in lieu of the actual data, to obtaining low-dimensional scalar equations that capture the behavior of the estimator, as we have demonstrated in simple examples of data augmentation. In fact, the majority of our analysis is not exclusive to logistic regression and can be directly extended to any classification algorithm such as SVM that relies on (X_i) solely through its one-dimensional projections $(X_i^T \beta)$ (see Section 8). Moreover, our dependent CGMT is not tied to the logistic model and relies only on a low-rank dependence assumption. An interesting future line of work would be to extend our analysis to high-dimensional models such as random feature models and to extend our dependency assumptions to a more general mixing condition (Bradley, 2005). As demonstrated in our plots, there is also no reason that universality should be a uniquely sub-Gaussian phenomenon, as opposed to a proof artifact. Extending this to other distributions would constitute a valuable extension of our work.

In our simulations, we also observed *non-universality of the training trajectories*. In Fig. 1 and 2, most estimates $\hat{\beta}$ are obtained via gradient descent with a learning rate 0.1 until either convergence or 10^6 steps are completed. Two exceptions are the t-distribution and the uniform distribution in Fig. 1: Numerically, we find that different learning rates are required to converge to the global minimum within 10^6 steps. Indeed, our results establish the universality of the *global minima*, but do not answer whether the *training trajectories to reach these minima* are universal, since the latter question is specific to the optimization methods employed. In Fig. 4 in the appendix, we observe that with the same learning rate, the training loss curves differ for uniform and t distributions, but agree for the remaining distributions. An interesting follow-up question to investigate is whether universality holds for training trajectories under different optimization methods.

Acknowledgments

MEM is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 2140743. KHH is supported by the Gatsby Charitable Foundation.

References

Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. *Advances in Neural Information Processing Systems*, 32, 2019.

Danil Akhtiamov, David Bosch, Reza Ghane, K Nithin Varma, and Babak Hassibi. A novel Gaussian min-max theorem and its applications. *arXiv preprint arXiv:2402.07356*, 2024a.

Danil Akhtiamov, Reza Ghane, and Babak Hassibi. Regularized linear regression for binary classification. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 202–207. IEEE, 2024b.

Sh A Alimov, RR Ashurov, and AK Pulatov. Multiple fourier series and fourier integrals. *Commutative Harmonic Analysis IV: Harmonic Analysis in IR n*, pages 1–95, 1992.

Liviu Aolaritei, Soroosh Shafieezadeh-Abadeh, and Florian Dörfler. The performance of Wasserstein distributionally robust M-estimators in high dimensions. *arXiv preprint arXiv:2206.13269*, 2022.

Morgane Austern and Peter Orbanz. Limit theorems for distributions invariant under groups of transformations. *The Annals of Statistics*, 50(4):1960–1991, 2022.

Sumanta Basu and George Michailidis. Regularized estimation in sparse high-dimensional time series models. 2015.

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*, 116(32):15849–15854, 2019.

Gregory Benton, Marc Finzi, Pavel Izmailov, and Andrew G Wilson. Learning invariances in neural networks from training data. *Advances in neural information processing systems*, 33:17605–17616, 2020.

Serge Bernstein. Sur l'extension du théorème limite du calcul des probabilités aux sommes de quantités dépendantes. *Mathematische Annalen*, 97:1–59, 1927.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 1995.

George Ebow Bonney. Logistic regression for dependent binary observations. *Biometrics*, pages 951–973, 1987.

Richard C Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–114, 2005.

William Brock, Josef Lakonishok, and Blake LeBaron. Simple technical trading rules and the stochastic properties of stock returns. *The Journal of finance*, 47(5):1731–1764, 1992.

Anthony Carbery and James Wright. Distributional and lq norm inequalities for polynomials over convex bodies in rn. *Mathematical research letters*, 8(3):233–248, 2001.

Sourav Chatterjee. *Concentration inequalities with exchangeable pairs*. Stanford University, 2005.

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.

Wei-Kuo Chen and Wai-Kit Lam. Universality of approximate message passing algorithms. *Electronic Journal of Probability*, 26:1 – 44, 2021.

Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 1993.

Jonathan D Cryer. *Time series analysis*, volume 286. Duxbury Press Boston, 1986.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for Gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36, 2024.

James Davidson. A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes. *Econometric theory*, 8(3):313–329, 1992.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2): 435–495, 2022.

Chandrakant M Deo. A note on empirical processes of strong-mixing sequences. *The Annals of Probability*, pages 870–875, 1973.

Oussama Dhifallah and Yue Lu. On the inherent regularization effects of noise injection during training. In *International Conference on Machine Learning*, pages 2665–2675. PMLR, 2021.

David L Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Physical Review E*, 109(3):034305, 2024.

Yehoram Gordon. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, 50:265–289, 1985.

Shaojun Guo, Yazhen Wang, and Qiwei Yao. High dimensional and banded vector autoregressions, 2016. URL https://arxiv.org/abs/1502.07831.

Qiyang Han and Yandi Shen. Universality of regularized regression estimators in high dimensions. *The Annals of Statistics*, 51(4):1799–1823, 2023.

Boris Hanin and Yi Sun. Data augmentation as stochastic optimization. 2020.

JK Haseman and LL Kupper. Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, pages 281–293, 1979.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949, 2022.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2022.

Kevin Han Huang, Peter Orbanz, and Morgane Austern. Data augmentation in the underparameterized and overparameterized regimes. *arXiv preprint arXiv:2202.09134*, 2022.

Kevin Han Huang, Xing Liu, Andrew Duncan, and Axel Gandy. A high-dimensional convergence theorem for u-statistics with applications to kernel-based testing. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3827–3918. PMLR, 2023.

I Ibragimov. Independent and stationary sequences of random variables. *Wolters, Noordhoff Pub.*, 1975.

Dunham Jackson. *The theory of approximation*, volume 11. American Mathematical Soc., 1930.

Adel Javanmard and Mahdi Soltanolkotabi. Precise statistical analysis of classification accuracies for adversarial training. *The Annals of Statistics*, 50(4):2127–2156, 2022.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results, 2013.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

Samriddha Lahiry and Pragya Sur. Universality in block dependent linear models with applications to nonparametric regression. *IEEE Transactions on Information Theory*, 70(12):8975–9000, 2024.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 1991.

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.

Clare Lyle, Mark van der Wilk, Marta Kwiatkowska, Yarin Gal, and Benjamin Bloem-Reddy. On the benefits of invariance in neural networks. In *Conference on Neural Information Processing Systems: Workshop on Machine Learning with Guarantees*, 2019.

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.

Francesca Mignacco, Florent Krzakala, Yue Lu, Pierfrancesco Urbani, and Lenka Zdeborova. The role of regularization in classification of high-dimensional noisy Gaussian mixture. In *International conference on machine learning*, pages 6874–6883. PMLR, 2020.

Andrea Montanari and Phan-Minh Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Andrea Montanari, Feng Ruan, Basil Saeed, and Youngtak Sohn. Universality of max-margin classifiers. *arXiv preprint arXiv:2310.00176*, 2023.

Somabha Mukherjee, Ziang Niu, Sagnik Halder, Bhaswar B Bhattacharya, and George Michailidis. High dimensional logistic regression under network dependence. *arXiv preprint arXiv:2110.03200*, 2021.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. *Advances in neural information processing systems*, 33:16666–16676, 2020.

William B. Nicholson, Ines Wilms, Jacob Bien, and David S. Matteson. High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research*, 21(166): 1–52, 2020. URL http://jmlr.org/papers/v21/19-777.html.

Manfred Opper, Ole Winther, et al. From naive mean field theory to the tap equations. *Advanced mean field methods: theory and practice*, pages 7–20, 2001.

Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

Joost D. J. Plate, Rutger R. van de Leur, Luke P.H. Leenen, Falco Hietbrink, Linda M. Peelen, and Marinus J. C. Eijkemans. Incorporating repeated measurements into prediction models in the critical care setting: a framework, systematic review and meta-analysis. *BMC Medical Research Methodology*, 19, 2019.

Ross L Prentice. Correlated binary regression with covariates specific to each binary observation. *Biometrics*, pages 1033–1048, 1988.

Beth A Reboussin, Edward H Ip, and Mark Wolfson. Locally dependent latent class models with covariates: an application to under-age drinking in the usa. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(4):877–897, 2008.

Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton Mathematical Series. Princeton University Press, Princeton, N. J., 1970.

Nathan Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210 – 293, 2011.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration, 2013.

Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Michael Schweinberger and Mark S Handcock. Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(3):647–676, 2015.

Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. 2019.

C. Shorten and T. M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

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

Mihailo Stojnic. A framework to characterize performance of lasso algorithms. *arXiv preprint arXiv:1303.7291*, 2013a.

Mihailo Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013b.

Pragya Sur and Emmanuel J Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, 2019.

Arwa Mohammed Taqi, Ahmed Awad, Fadwa Al-Azzo, and Mariofanna Milanova. The impact of multi-optimizers and data augmentation on tensorflow convolutional neural network performance. In *Proc. of IEEE MIPR*, pages 140–145, 2018.

Christos Thrampoulidis. *Recovering structured signals in high dimensions via non-smooth convex optimization: Precise performance analysis*. PhD thesis, California Institute of Technology, 2016.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The Gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized *m*-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Ruey S Tsay. *Analysis of financial time series*. John wiley & sons, 2005.

D Tuan and T Lanh. Some mixing properties of time series models. *Stochastic processes and their applications*, 19:297–303, 1985.

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

Svetlana Volkova. An overview on data augmentation for machine learning. In Arthur Gibadullin, editor, *Digital and Information Technologies in Economics and Management*, pages 143–154, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-55349-3.

Hans Wackernagel. *Multivariate geostatistics: an introduction with applications*. Springer Science & Business Media, 2003.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *The Annals of Applied Probability*, 34(4):3943–3994, 2024.

Kam Chung Wong, Ambuj Tewari, and Zifan Li. Regularized estimation in high dimensional time series under mixing conditions. *stat*, 1050:12, 2016.

Kam Chung Wong, Zifan Li, and Ambuj Tewari. Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, 48(2):1124–1142, 2020.

Margaret Wu and James H Ware. On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics*, pages 513–521, 1979.

Jianke Yang, Robin Walters, Nima Dehmamy, and Rose Yu. Generative adversarial symmetry discovery. In *International Conference on Machine Learning*, pages 39488–39508. PMLR, 2023.

Bin Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, pages 94–116, 1994.

Guo Yu and Jacob Bien. Learning local dependence in ordered data. *Journal of Machine Learning Research*, 18(42):1–60, 2017.

Christopher JW Zorn. Generalized estimating equation models for correlated data: A review with applications. *American Journal of Political Science*, pages 470–490, 2001.

Bin Zou, Luoqing Li, and Zongben Xu. The generalization performance of erm algorithm with strongly mixing observations. *Machine learning*, 75(3):275–295, 2009.

The appendix is organized as follows:

- Appendix A presents additional definitions and notation used throughout the appendix.
- Appendix B state additional results. This includes the characterizing (EQs) for selected data augmentation in Appendix B.1, a comparison of our full dependent CGMT to the classical CGMT in Appendix B.2, a generalized logistic model in Appendix B.3, and a brief discussion on Assumption 10 in Appendix B.4.
- Appendix C includes simulation details.
- Appendix D proves training risk universality in Theorem 2 for the more general model in Appendix B.3 under block dependence. The generalized result is stated in Appendix D.2. Note that in Appendices D and H, we temporarily convert the $0/1$ labels to ± 1 as it simplifies the proofs; the equivalence between the two label schemes is performed in Appendix D.3.
- Appendix E proves training risk universality in Theorem 3 under Assumption 8(i) for m-dependent processes.
- Appendix F proves training risk universality in Theorem 3 under Assumption 8(ii) for mixing processes.
- Appendix G proves test risk universality in Theorems 2 and 3.
- Appendix H collects important lemmas used for the proofs in Appendices D and G.
- Appendix I contains a result that establishes the equivalence of training losses under deletion of small blocks, which is utilised in the big-block-small-block technique applied in Appendices E and F.
- Appendix J verifies Assumption 5 for different augmentation schemes.
- Appendix K collects auxiliary lemmas used for the proofs in Appendices G and J.
- Appendix L proves the dependent CGMT.
- Appendix M present all intermediate optimizations used for applying CGMT to analyze data augmentation. We also include results that verify the CGMT conditions for different augmentations.
- Appendix N proves all results in Appendix M.

Appendix A. Additional Definitions and Notations

Our results hold under the assumption that the random vectors (X_i) are sub-Gaussian. We present here a formal definition of sub-Gaussianity.

Definition 9 *We say that a random vector $Y \in \mathbb{R}^p$ is sub-Gaussian with constant K if, for all vectors $\boldsymbol{v} \in \mathbb{R}^p$, we have*

$$\mathbb{E}\left[\exp\left(\lambda\langle \boldsymbol{v}, Y - \mathbb{E}(Y)\rangle\right)\right] \leq \exp\left(C\lambda^2\|\boldsymbol{v}\|_2^2\mathsf{K}^2\right) \qquad \textit{for all } \lambda > 0,$$

for some absolute constant $C > 0$.

If Y is sub-Gaussian, then the norm of its covariance matrix is well controlled (see Lemma 44 for more details). Furthermore, a number of results assume that the data is locally dependent, as defined in Ross (2011).

Definition 10 *Let $(X_i)_{i\leq p} \in \mathbb{R}^p$ be a random vector. We say that it is locally dependent if for all $i \leq p$ there exists a subset $\mathcal{N}_i \subset [p]$ such that X_i is independent from $(X_k)_{k\notin \mathcal{N}_i}$. We call $\mathcal{N}_i$ the dependency neighborhood of X_i.*

A similar definition can be made for random arrays:

Definition 11 *Let $(X_{i,j})_{i\leq p_1, j\leq p_2}$ be a random array. We say that it is locally dependent if for all $i \leq p_1$ and $j \leq p_2$ there exists a subset $\mathcal{N}_{i,j} \subset [p_1] \times [p_2]$ such that $X_{i,j}$ is independent from $(X_{k,l})_{(k,l)\notin \mathcal{N}_{i,j}}$.*

Throughout the appendix we use the following notation:

- For a sequence (W_i) and a set $B \subset \mathbb{N}$, we let W_B designate $(W_i)_{i\in B}$.
- Recall the definition of the blocks $\mathcal{B}_i$ in (1). In our block dependent proofs, we may assume $\mathcal{B}_i = \{i, i+1, \ldots, i+k-1\}$ for notational simplicity, without any loss of generality.
- For a matrix $A \in \mathbb{R}^{n\times n}$, $a_{\mathcal{B}_i}$ denotes $(a_{ij})_{j\in\mathcal{B}_i} \in \mathbb{R}^k$.
- For a set $\mathcal{S}$ and $\delta > 0$, we let $\mathcal{S}_\delta$ designate a minimal $\delta\sqrt{p}$-net of $\mathcal{S}$.
- We use $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$.
- The function $\mathbf{1}^{\pm}(x)$ will be used to denote the sign function $\mathrm{sgn}(x) = \mathbb{I}(x \geq 0) - \mathbb{I}(x < 0)$.
- Any constant in sans serif font such as C_1, C_2, and so on, depends on at most the constants L, K_X, and κ given in our assumptions. If it further depends on δ for example, then it will be written as C_δ or $\mathsf{C}(\delta)$.
- We write $\hat{\beta}(\mathbf{X})$ as simply $\hat{\beta}$ when it is clear from context.
- For a matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$ and a given set $\mathcal{S} \subseteq \mathbb{R}^p$, the "scaled" operator norm on $\mathcal{S}$ is defined via

$$\|\boldsymbol{Y}\|_{\mathcal{S}} := \sup_{\beta\in\mathcal{S}} \|\mathbf{Y}\beta\|. \tag{12}$$

Appendix B. Additional Results

B.1. Characterizing equations for selected data augmentations

As discussed in Section 5, the dependent CGMT allows us to derive explicitly a set of deterministic, low-dimensional equations that capture the asymptotic behavior of a logistic regression under data augmentations. As an example, we compute this explicitly under a further simplifying assumption on the covariance structure of the augmented data. To state the assumption, let $\Sigma_o := \mathrm{Var}[Z_1] \in \mathbb{R}^{p\times p}$, $\Sigma := \mathrm{Var}[X_1] = \mathrm{Var}[\phi_1(Z_1)] \in \mathbb{R}^{p\times p}$, and $\Sigma_o^{\dagger}, \Sigma^{\dagger}$ be their respective pseudo-inverses.

Assumption 11 *Write $\Sigma_* := (\Sigma^{\dagger})^{1/2}\,\mathrm{Cov}[\phi_1(Z_1)\,,\,\phi_2(Z_1)]\,(\Sigma^{\dagger})^{1/2}$. Assume that*

$$(i)\ \Sigma_* \;= (\Sigma^{\dagger})^{1/2}\,\mathrm{Cov}[\phi_1(Z_1)\,,\,Z_1](\Sigma_o^{\dagger})^{1/2} \qquad \textit{and} \qquad (ii)\ \Sigma_*^2 \;= \Sigma_*\,.$$

Since ϕ_1 and ϕ_2 are i.i.d. transformations, Assumption 11(i) holds for example under an invariance assumption, $Z_1 \overset{d}{=} \phi_1(Z_1)$. Assumption 11(ii) requires that the eigenvalues of Σ_* consist of only zeros and ones. Note that Σ_* is symmetric and idempotent, and therefore a projection matrix; this property is exploited throughout the CGMT formula computation in Appendix L.

We verify Assumption 11 for the cases of no augmentation, random permutation and random sign flipping in Appendix M.2. Note that Assumption 11 is more restrictive than necessary: While it does not cover random cropping, the CGMT theorem does apply to random cropping and the only difference is that one cannot use Assumption 11 to simplify certain algebraic calculations, resulting in a more complicated set of equations than (EQs). We clarify this in Appendix M.2.

To apply the CGMT to obtain a deterministic set of equations, one needs to establish the equivalence of multiple optimization problems. We state them in Appendix M, which includes the original optimization (OO) (i.e. (11) in Section 6), the primary optimization (PO) (i.e. $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$ in Theorem 5), the auxiliary optimization (AO) (i.e. $\psi_{\mathcal{S}_w,\mathcal{S}_u}$ in Theorem 5), a low-dimensional scalar optimization (SO) and a low-dimensional deterministic optimization (DO), whose solutions are characterized by (EQs). We state the main result here.

Theorem 12 (Effect of data augmentation on the test risk) *Let $\hat{\beta}(\mathbf{X}, \mathbf{X}^{\Phi})$ be the estimator fitted via* (OO) *with $S = \mathcal{S}_p$. Assume that Assumptions 1 – 6 hold, that the minimizer-maximizers of* (DO) *are within the interior of the domain of optimization and Assumption 11 holds. Then*

$$|R_{\text{test}}(\hat{\beta}(\mathbf{X}, \mathbf{X}^{\Phi})) - R^{G}_{\text{test}}(\hat{\beta}(\mathbf{G}, \mathbf{G}^{\Phi}))| \xrightarrow{\mathbb{P}} 0 \qquad \textit{and} \qquad |R_{\text{test}}(\hat{\beta}(\mathbf{X}, \mathbf{X}^{\Phi})) - \bar{R}_{\text{test}}(\bar{\chi}_2^{r,\theta,\sigma,\tau})| \xrightarrow{\mathbb{P}} 0 \,,$$

where $(r, \theta, \sigma, \tau)$ solves the system of equations (EQs)*, $\bar{\chi}_2^{r,\theta,\sigma,\tau}$ is defined in* (DO)*, and*

$$\bar{R}_{\text{test}}(\bar{\chi}) \;:=\; \mathbb{E}_{\eta\sim\mathcal{N}(0,1)}[\ell_{\text{test}}(\, \sqrt{\bar{\chi}}\,\eta\,,\, \|\Sigma^{1/2}_{\text{new}}\beta^*\|\eta)] \,.$$

Theorem 12 shows that the test risk is completely characterized by a 1d quantity $\bar{\chi}_2^{r,\theta,\sigma,\tau}$. This quantity is completely determined by the parameters $(\alpha, \sigma_1, \sigma_2, \tau_1, \tau_2, \nu_1, \nu_2, r_1, r_2, \theta)$, defined as solutions to the system of 10 non-linear equations

$$\begin{cases}
0 \;=\; \theta\bar{\kappa}_*^2 - \frac{\alpha\bar{\kappa}_*^2}{\sigma_2\tau_2} - \frac{r_2\nu_2\bar{\kappa}_*}{k}\mathbb{E}\left[\bar{Z}_1\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta}\right] + r_2\nu_2\alpha\bar{\kappa}_*^2 \,, \\
0 \;=\; -\frac{1}{2\tau_1} - \partial_{\sigma_1}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \frac{r_1\nu_1}{k}\,\mathbb{E}\Big[\eta^\intercal\,(I_k - \frac{1}{k}\mathbf{1}_{k\times k})\,u_{\bar{Z},\varepsilon_1,\eta}\Big] + \frac{r_1\nu_1\sigma_1(k-1)}{k} + \frac{r_2\nu_2}{k}\mathbb{E}\Big[\eta^\intercal\frac{1}{k}\mathbf{1}_{k\times k}u_{\bar{Z},\varepsilon_1,\eta}\Big] \\
\qquad + \frac{r_2\nu_2\sigma_1}{k} \,, \\
0 \;=\; -\frac{1}{2\tau_2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2^2\tau_2} - \partial_{\sigma_2}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \frac{r_2\nu_2}{k}\mathbb{E}\Big[\bar{Z}_2\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta}\Big] + r_2\nu_2\sigma_2 \,, \\
0 \;=\; \frac{\sigma_1}{2\tau_1^2} - \partial_{\tau_1}\bar{\chi}_1^{r,\theta,\sigma,\tau} \,, \\
0 \;=\; \frac{\sigma_2}{2\tau_2^2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2^2} - \partial_{\tau_2}\bar{\chi}_1^{r,\theta,\sigma,\tau} \,, \\
0 \;=\; -\frac{r_1}{2\nu_1^2} + \frac{r_1}{2k}\,\mathbb{E}\Big[\Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1\eta)\Big\|^2\Big] \,, \\
0 \;=\; -\frac{r_2}{2\nu_2^2} + \frac{1}{4r_2\nu_2^2} + \frac{r_2}{2k}\mathbb{E}\Big[\Big\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\Big\|^2\Big] \\
\qquad + \frac{1}{\nu_2 k}\mathbb{E}\Big[\bar{Y}\Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2\nu_2}\bar{Y} - k\alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \,, \\
0 \;=\; \frac{1}{2\nu_1} - \partial_{r_1}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \frac{\nu_1}{2k}\,\mathbb{E}\Big[\Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1\eta)\Big\|^2\Big] \,, \\
0 \;=\; \frac{1}{2\nu_2} + \frac{1}{4r_2^2\nu_2} - \partial_{r_2}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \frac{\nu_2}{2k}\mathbb{E}\Big[\Big\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\Big\|^2\Big] \\
\qquad + \frac{1}{r_2 k}\mathbb{E}\Big[\bar{Y}\Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2\nu_2}\bar{Y} - k\alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \,, \\
0 \;=\; \alpha\bar{\kappa}_*^2 - \partial_{\theta}\bar{\chi}_1^{r,\theta,\sigma,\tau} \,.
\end{cases} \tag{EQs}$$

Here, $\bar{Z} = (\bar{Z}_0, \bar{Z}_1, \bar{Z}_2)$ and $\eta = (\eta_1, \ldots, \eta_k)$ are both independent low-dimensional standard Gaussians, ε_1 is an independent Logistic-(0, 1) variable, $\bar{Y} := \mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\} = \mathbb{I}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1 \geq 0\}$, and $\bar{\kappa}_*$, $\bar{\kappa}_0$ and $\bar{\chi}_1^{r,\theta,\sigma,\tau}$ are limits defined in (DO) that are related to β^* and the covariances of the original data as well as the augmented data. $u_{\bar{Z},\varepsilon_1,\eta}$ can be viewed as a generalization of the proximal operator used in Salehi et al. (2019), in the sense that its is defined as an minimizer of the low-dimensional, random optimization problem

$$\begin{aligned}\min_{\tilde{u}\in\mathbb{R}^k} \frac{1}{k}\mathbf{1}_k^\top\rho(\tilde{u}) + \frac{r_1\nu_1}{2k}\left\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(\tilde{u}+\sigma_1\eta)\right\|^2 \\ + \frac{r_2\nu_2}{2k}\left\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(\tilde{u} - \frac{1}{r_2\nu_2}\bar{Y}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\right\|^2 .\end{aligned} \tag{13}$$

While the system of equations is rather complicated, we show in Lemma 52 that it recovers exactly the system of 6 non-linear equations in the case of isotropic data with no augmentation, derived in Salehi et al. (2019). As part of our proof, we also observe that σ_1, τ_1, ν_1 and r_1 are the additional parameters that arise due to augmentation.

Proof of Theorem 12 By Theorem 15, the conclusion of Theorem 2 (5) holds for the logistic model (12). This in particular includes the data augmentation model (11) by considering an $m(k+1)$ dataset $(Z_{i'}, \phi_{i'(k-1)+1}(Z_{i'}), \dots, \phi_{i'k}(Z_{i'}))_{i'\leq m}$, setting the weights ω_i of the loss to be 1 for all augmented data and 0 for the unaugmented data, and setting the weights a_{ij} in the labels such that the labels of $\phi_{i'(k-1)+1}(Z_{i'}), \dots, \phi_{i'k}(Z_{i'})$ all depend only on $Z_{i'}$. Meanwhile, notice that the proof of Theorem 2 (6) in Appendix G does not depend on the choice of the logistic model as long as training risk universality is estbalished. Therefore the test risk universality in Theorem 2 (6) would hold for data augmentation and the stated assumptions, if Assumption 7 is verified.

To verify Assumption 7, we set $\bar{\chi} = \bar{\chi}_2^{r,\theta,\sigma,\tau}$ in Assumption 7, and combine Lemma 47, Lemma 49 and Lemma 50 to relate (GO) to (DO). By assumption, the minimizer-maximizers of (DO) are within the interior of the domain of optimization, so the converged risk of (DO) changes by $\Theta(\epsilon^2)$ depending on whether the optimization domain of β requires $|(\beta^\top\Sigma_{\text{new}}\beta)^{1/2} - (\bar{\chi}_2^{r,\theta,\sigma,\tau})^{1/2}| > \epsilon$. This verifies Assumption 7 and proves the universality of the test risk. The deterministic approximation then follows by substituting $\bar{\chi} = (\bar{\chi}_2^{r,\theta,\sigma,\tau})^{1/2}$ and applying Lemma 51 to obtain (EQs). ■

B.2. Dependent and classical CGMT results

The next result states the full version of our dependent CGMT, for which Theorem 5 is a direct corollary. $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$ and $\psi_{\mathcal{S}_w,\mathcal{S}_u}$ are the risks under the primary optimization (9) and the auxiliary optimization (10) respectively, both defined in Section 5; $\hat{w}_\Psi \in \mathcal{S}_w$ is the minimizer of $\Psi_{\mathcal{S}_w,\mathcal{S}_u}$.

Theorem 13 (Dependent CGMT) *Suppose $\mathcal{S}_w$ and $\mathcal{S}_u$ are compact and f is continuous on $\mathcal{S}_w \times \mathcal{S}_u$. Then the following statements hold:*

(i) For all $c \in \mathbb{R}$,

$$\mathbb{P}(\Psi_{\mathcal{S}_w,\mathcal{S}_u} \leq c) \leq 2^M\mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \leq c) .$$

(ii) If additionally $\mathcal{S}_w$ and $\mathcal{S}_u$ are convex and f is convex-concave on $\mathcal{S}_w \times \mathcal{S}_u$, then for all $c \in \mathbb{R}$,

$$\mathbb{P}(\Psi_{\mathcal{S}_w,\mathcal{S}_u} \geq c) \leq 2^M\mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \geq c) ,$$

and in particular, for all $\mu \in \mathbb{R}$ and $t > 0$,

$$\mathbb{P}(|\Psi_{\mathcal{S}_w,\mathcal{S}_u} - \mu| \geq t) \leq 2^M\,\mathbb{P}(|\psi_{\mathcal{S}_w,\mathcal{S}_u} - \mu| \geq t) .$$

(iii) *Assume the conditions of (ii). Let $\mathcal{A}_p$ be an arbitrary open subset of $\mathcal{S}_w$ and $A_p^c := \mathcal{S}_w \setminus \mathcal{A}_p$. If there exists constants $\bar{\psi}_{\mathcal{S}_w}$, $\bar{\psi}_{\mathcal{A}_p^c}$ and $\eta, \epsilon > 0$ such that*

$$\bar{\psi}_{\mathcal{A}_p^c} \geq \bar{\psi}_{\mathcal{S}_w} + 3\eta\ , \quad \mathbb{P}(\psi_{\mathcal{S}_w,\mathcal{S}_u} \leq \bar{\psi}_{\mathcal{S}_w} + \eta) \geq 1 - \epsilon\ , \quad \mathbb{P}(\psi_{\mathcal{A}_p^c,\mathcal{S}_u} \geq \bar{\psi}_{\mathcal{A}_p^c} - \eta) \geq 1 - \epsilon\ ,$$

then $\mathbb{P}(\hat{w}_\Psi \in \mathcal{A}_p) \geq 1 - 4\epsilon$.

As a comparison, we remark that the standard CGMT in the isotropic, independent case is exactly the same as above with $M = 1$, and stated for the loss

$$\begin{aligned}\Psi_{\mathcal{S}_w,\mathcal{S}_u} &:= \min_{w\in\mathcal{S}_w} \max_{u\in\mathcal{S}_u} w^\top \mathbf{H} u + f(w,u)\\ \tilde{\psi}_{\mathcal{S}_w,\mathcal{S}_u} &:= \min_{w\in\mathcal{S}_w} \max_{u\in\mathcal{S}_u} \|w\|_2 \mathbf{h}^\top u + w^\top \mathbf{g}\|u\| + f(w,u)\ .\end{aligned}$$

In this case, $\mathbf{H}$ is an $\mathbb{R}^{p\times n}$ matrix with i.i.d. standard Gaussian entries, and $\mathbf{h}$ and $\mathbf{g}$ are again independent standard Gaussian vectors in $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively. We refer interested readers to Thrampoulidis (2016) for a detailed overview of CGMT and their Theorem 3.3.1 for the standard CGMT result.

B.3. Generalizing the Model

Recall that the model stated in Section 1.1 assumes that y_i is only a function of its own covariates:

$$\mathbb{P}(y_i = 1 \mid X_i) = \sigma\left(X_i^\top \beta^*\right).$$

However, this formulation is quite limiting with regards to the types of dependence it can handle. Recall that a key property of data augmentation, for example, is that any transformation we apply to the covariates should not alter the associated label (meaning $y_1 = y_2 = \cdots = y_k$). This suggests that our model must be able to account for both the classical setup of logistic regression, and that of data augmentation, repeated measurements, and more. Thus, for the rest of our block-dependent results and proofs given in the appendix, we alter Assumption 2 in the following way:

Assumption 12 (Generalized Model) *There exists a block diagonal matrix $A = (a_{ij}) \in [0,1]^{n\times n}$ satisfying $a_{ij} = 0$ if $j \notin \mathcal{B}_i$ and $\sum_{j\in\mathcal{B}_i} a_{ij} = 1$ for all $1 \leq i \leq n$, such that*

$$\mathbb{P}\left(y_i = 1\right) = \sigma\Big(\textstyle\sum_{j\in\mathcal{B}_i} a_{ij} X_j^\top \beta^*\Big).$$

Recalling that our training risk is given by

$$\hat{R}_n(\beta, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^n \omega_i \left(\log\left(1 + e^{X_i^\top \beta}\right) - y_i X_i^\top \beta\right) + \frac{\lambda}{2n}\|\beta\|^2,$$

we can specify certain values of our matrix A and the weights $\omega := (\omega_1, \ldots, \omega_n)$ to obtain relevant setups:

(i) When $A = I_n$ and $\omega = (1, \ldots, 1)$, we recover the classic logistic regression framework.

(ii) When $a_{ij} = \mathbb{I}(j = \min \mathcal{B}_i)$ and $\omega_i = \mathbb{I}(i \neq \min \mathcal{B}_i)$, we obtain the data augmentation framework as utilized in Hu and Lu (2022). Note that this implies that the first element of each block is the original data point which defines the labels, but is not considered in the regression.

(iii) If $(a_{ij})_{j\in\mathcal{B}_i} = (\frac{1}{k}, \ldots, \frac{1}{k})$, the label is defined by an equally weighted sum of the block, which can be utilized for situations such as repeated measurements and peer effects.

B.4. Discussion on Assumption 10

The low-rank covariance structure in Assumption 10 is natural for setups with data augmentation, as illustrated by the next lemma.

Lemma 14 *Fix m such that m divides n and write $k := m/n$. Let $X_1, \ldots, X_m$ be i.i.d. random vectors in $\mathbb{R}^p$, let $\phi_1, \ldots, \phi_n$ be i.i.d. random $\mathbb{R}^d \to \mathbb{R}^d$ transformations, and let $\mathbf{G}$ be an $\mathbb{R}^{p\times n}$-valued Gaussian matrix that matches the mean and covariance of the augmented data matrix*

$$(\phi_1(X_1), \ldots, \phi_k(X_1), \ldots, \phi_{(m-1)k+1}(X_m), \ldots, \phi_{mk}(X_m)) \; .$$

Also denote $\Sigma'_1 := \mathrm{Var}[\phi_1(X_1)]$ and $\Sigma'_2 := \mathrm{Cov}[\phi_1(X_1), \phi_2(X_1)]$. Then $\mathbf{G}$ satisfies Assumption 10 with

$$\mathrm{Cov}[\mathbf{G}_{ji}, \mathbf{G}_{j'i'}] \;=\; (\Sigma_1 - \Sigma_2)\mathbb{I}\{i = i'\} + \Sigma_2 \mathbb{I}\{i \in N(i')\} \qquad \textit{for all } i, i' \leq n \textit{ and } j, j' \leq p \; ,$$

where $N(i') := \{\lfloor (i'-1)/k \rfloor + 1, \ldots, \lfloor (i'-1)/k \rfloor + k\}$ is the set of indices that correspond to differently augmented versions of the same data vector. Moreover, $\Sigma_1 - \Sigma_2$ is positive semi-definite.

Proof of Lemma 14 The proof is identical to that of Lemma 58 in Appendix N by replacing $P_*^\perp$ with identity. ■

We do note that, however, in the actual application of CGMT, the Gaussian matrix considered in Assumption 10 is typically a slightly modified version of $\mathbf{H}$, although most of the dependence structure is typically inherited. In our examples, this modification comes from a projection matrix analogous to those used in the i.i.d. version of CGMT. A precise formulation is included in Appendix M, and the corresponding verification of Assumption 10 is presented in Lemma 58 of Appendix N.

Appendix C. Simulation Details

We present some additional simulation details on top of the setups described in Section 6 here. The regularization parameter is held at $\lambda = 0.01$, and the test loss is computed as the difference between the 0-1 loss achieved by $\hat{\beta}$ and that achieved by the oracle β^*. Below, we denote $\mathcal{N}$, Unif, Γ_2, Exp and t_3 respectively as a standard normal, a uniform distribution, a gamma distribution with shape 2, an exponential distribution and a Student's t distribution with 3 degrees of freedom, all shifted and rescaled to have zero mean and $1/p$ variance. We also write $\tilde{t}_3$ as t_3 rescaled to have unit variance.
Details for random permutations. Fig. 1 concerns the performance of random permutations across different proportion r_{perm} of coordinates to permute and different number of augmentations k. Results are collected over 50 random trials for augmented data and over 200 random trials for unaugmented data. The dimension is fixed as $p = 500$, 50 groups are considered, and all group sizes are kept the same with $p_1 = \ldots = p_{50} = 10$. In each trial, $\beta^* \in \mathbb{R}^p$ is generated by concatenating 50 groups of 10 identical entries each, where the 50 different entries are generated i.i.d. from $\tilde{t}_3$. Every group of coordinates of the data are generated i.i.d. according to $\mathcal{N}$, Unif, Γ_2, Exp and t_3, but additionally rescaled by a random group-dependent parameter drawn from $\Gamma(0.5, 1)$. The choices of $r_{\text{perm}} = 0.8$ and $r_{\text{perm}} = 1.0$ correspond to random permutations performed respectively on the top 8 and 10 coordinates of each group. Plots with no augmentation are generated under Gaussian data.

Details for random cropping and sign flipping. Fig. 2 concerns the performance of random cropping and random sign flipping across different signal ratio ρ^* and different data dimension p.

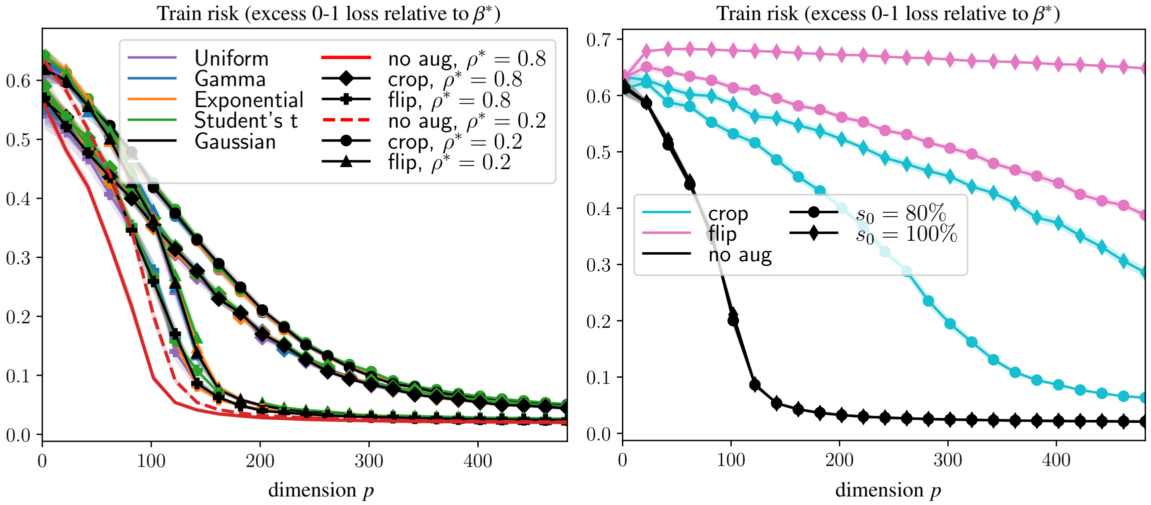

Figure 3: Universality of training risks under cropping and sign flipping. The left and the right plots are the training risk analogues of the left and the right plots of Fig. 2 respectively.

Results are collected over 100 random trials for augmented data and over 200 random trials for unaugmented data. The number of augmentations is fixed as $k = 30$. Plots with no augmentation are generated under Gaussian data.

- For the setup without knowledge of zero coordinates (left plot of Fig. 2 and left plot of Fig. 3), β^* is generated such that a uniformly random subset of $\lceil (1 - \rho^*)p \rceil$ coordinates are zero and the remaining entries are drawn i.i.d. from $\tilde{t}_3$, and random cropping and sign flipping are performed on $r_{\text{flip}} = r_{\text{crop}} = 20\%$ of the coordinates. Data are generated coordinate-wise i.i.d. according to $\mathcal{N}$, Unif, Γ_2, Exp and t_3.
- For the setup where the bottom $\lceil s_0(1 - \rho^*)p \rceil$ coordinates are known to be zero (right plot of Fig. 2 and right plot of Fig. 3), the remaining coordinates of β^* are generated such that a random subset of $\lceil (1 - \rho^*)p \rceil - \lceil (1 - \rho^*)p \rceil$ coordinates are zero and the rest are again drawn i.i.d. from $\tilde{t}_3$. Cropping and sign flipping are always performed on the bottom $\lceil s_0(1 - \rho^*)p \rceil$ coordinates, as well as also on $\lceil r \rceil - \lceil s_0(1 - \rho^*)p \rceil$ of the remaining coordinates, where $r = r_{\text{flip}} = r_{\text{crop}} = 0.2$. Data are generated coordinate-wise i.i.d. according to $\mathcal{N}$.

We also remark that even with knowledge of the coordinates, sign flipping does not outperform no augmentation: Unlike cropping, sign flipping does not explicitly leave out the zero coordinates.

Universality of risks. Notice that the simulations are performed over different distributions on the coordinates of Z_i's, shifted and scaled to have zero mean and the same variance. Notably, the uniform distribution obeys the sub-Gaussianity in Assumption 3, the exponential and gamma distributions only satisfy sub-exponential tails, and the t-distribution is chosen with 3 degrees of freedom, i.e. with unbounded third moments. Universality behavior is observed across all distributions. Indeed in our proof, sub-Gaussianity is only assumed for convenience, and we conjecture that this is not a necessary assumption for Theorem 5.

Non-universality of training trajectories as observed by the requirement of different learning rates. In both Fig. 1 and 2, gradient descent is employed to optimize the logistic regressor either until convergence or until 10^6 steps are exhausted. Learning rate is chosen as LR $= 0.1$ across all simulations with three exceptions: LR $= 1$ for t_3 in Fig. 1 under $r_{\text{perm}} = 0.8$, LR $= 0.5$ for

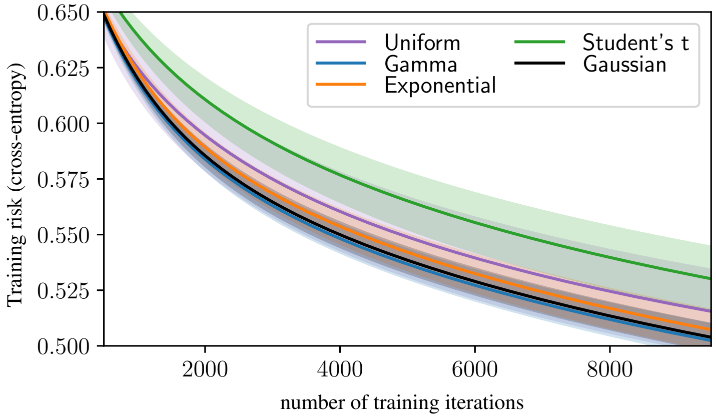

Figure 4: Initial training loss curves for the random permutation setup in Fig. 1 with $\rho_{\text{perm}} = 0.8$, $k = 11$ and learning rate $\text{LR} = 0.1$.

uniform distribution in Fig. 1 under $r_{\text{perm}} = 1.0$ and $\text{LR} = 0.8$ for uniform distribution in Fig. 1 under $r_{\text{perm}} = 0.8$. We find that for these three setups, $\text{LR} = 0.1$ does not lead to convergence within 10^5 steps. We conjecture that this arises due to the lack of universality of the training trajectories, as illustrated in Figure 4 and as discussed towards the end of Section 6.

Appendix D. Proof of Theorem 2 (5): Training risk universality

D.1. Restricting to $\mathcal{S}_p$

Recall that in Equation (4) we defined the set we are taking our minimum over as

$$\mathcal{S}_p := \left\{\beta \in \mathbb{R}^p : \|\beta\|_2 \leq \mathsf{L}\sqrt{p},\ \|\beta\|_\infty \leq \mathsf{L}p^{\frac{1-r}{2}}\right\}.$$

This restriction is very commonly applied in many results on universality, such as Lahiry and Sur (2024); Han and Shen (2023); Montanari and Saeed (2022). In our case, the Euclidean norm is trivial, since by virtue of being a minimizer we know that

$$\begin{aligned}\hat{R}_n(\hat{\beta}) \leq \hat{R}_n(0) &\implies \frac{1}{n}\sum_{i=1}^n \omega_i \left(\log\left(1 + e^{X_i^\top \hat{\beta}}\right) - y_i X_i^\top \hat{\beta}\right) + \frac{\lambda}{2n}\|\hat{\beta}\|^2 \leq \log(2) \\ &\implies \|\hat{\beta}\| \leq \sqrt{\frac{\log(4)}{\lambda}}\sqrt{n}.\end{aligned}$$

However, establishing the infinity norm bound with high probability is much more challenging. Such a bound is often proven through a leave-one-out approach, as in Karoui (2013); Han and Shen (2023), in which one defines a new minimizer

$$\hat{\beta}^{(s)} := \underset{\substack{\beta\in\mathbb{R}^p \\ \beta_s=0}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^n \log\left(1 + e^{-y_i X_i^\top \beta}\right) + \frac{\lambda}{2n}\|\beta\|_2^2$$

for each $1 \leq s \leq p$. For a well-behaved design, it should be the case that $\|\hat{\beta} - \hat{\beta}^{(s)}\|$ is small, which leads to a bound on $\|\hat{\beta}\|_\infty$. This idea was even generalized in Lahiry and Sur (2024) for block dependence within observations, where the new minimizer assumes that $\beta_{\mathcal{B}_i} = 0$ for an entire block $\mathcal{B}_i$.

Since our dependence is also across observations, these approaches are not able to leverage the independence between coordinates which plays a crucial role in such bounds. A more viable approach is taken in Theorem 5 & Lemma 13 of Montanari and Saeed (2022): they prove that under certain conditions, if a minimizer with bounded ℓ_2 norm exists (with high probability), then there also exists a minimizer with both bounded ℓ_2 and ℓ_∞ norms (with high probability). The only drawback to such an approach is that it requires a lower bound on the smallest singular value of a matrix, of the form

$$\sigma_{\min}(\mathbf{X}\mathbf{X}^\intercal) \geq \frac{p}{\mathsf{C}}.$$

This can be nearly impossible in some dependent set-ups: in the worst-case scenario, where all the rows in a neighborhood are identical, we know $\sigma_{\min}(\mathbf{X})$ is identically zero and such a bound is unattainable. However, certain DA schemes possess a nice enough structure to make this result hold and thus also the ℓ_∞ bound. In noise injection, for example, we can decompose the augmented matrix into a matrix of identical rows and the independent Gaussian noise. This latter matrix is asymptotically free from the first one, and thus we can obtain control on its smallest singular value. To extend this into a proof covering all DA schemes will be left to future work, and for now we will work under the mild constraint that $\hat{\beta} \in \mathcal{S}_p$.

D.2. Generalized Theorem

In this section, we will prove (5) from Theorem 2. However, we actually prove a slightly more general result:

Theorem 15 *Let $(X_i, y_i(X_i))_{i=1}^n$ and $(G_i, y_i(G_i))_{i=1}^n$ be generated under Assumptions 1, 3-5, and 12, where each $G_i \sim \mathcal{N}(0, \mathrm{Var}(X_i))$. Then for any $\tilde{\mathcal{S}} \subseteq \mathcal{S}_p$,*

$$d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n(\beta; \mathbf{X}), \min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n(\beta; \mathbf{G})\right) \to 0.$$

We remark that if we successfully establish Theorem 15, then the claim in (5) of Theorem 2 directly follows by setting $\tilde{\mathcal{S}} = \mathcal{S}_p$.

D.3. Converting the Loss

Before continuing, we will convert our labels from $\{0, 1\}$ to $\{-1, 1\}$, as this combines two of the terms in the training risk to significantly simplify calculations. To be specific, noting that $y_i \in \{0, 1\}$, we can define $\tilde{y}_i := 2y_i - 1 \in \{-1, 1\}$, which still satisfies

$$\mathbb{P}(\tilde{y}_i = 1 \mid X_i) = \mathbb{P}(y_i = 1 \mid X_i) = \sigma(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^*).$$

Then the loss evaluated at each data point (X_i, y_i) can be re-expressed as

$$\begin{aligned}
\log(1 + e^{X_i^\intercal\beta}) - y_i X_i^\intercal\beta &= \left(\log(1 + e^{X_i^\intercal\beta}) - X_i^\intercal\beta\right)\mathbb{I}\{y_i = 1\} + \log(1 + e^{X_i^\intercal\beta})\,\mathbb{I}\{y_i = 0\} \\
&= \log\left(\frac{1 + e^{X_i^\intercal\beta}}{e^{X_i^\intercal\beta}}\right)\mathbb{I}\{y_i = 1\} + \log(1 + e^{X_i^\intercal\beta})\,\mathbb{I}\{y_i = 0\} \\
&= \log(1 + e^{-X_i^\intercal\beta})\,\mathbb{I}\{\tilde{y}_i = 1\} + \log(1 + e^{X_i^\intercal\beta})\,\mathbb{I}\{\tilde{y}_i = -1\} \\
&= \log(1 + e^{-\tilde{y}_i X_i^\intercal\beta}).
\end{aligned}$$

Thus, renaming our labels as $y_i \in \{-1, 1\}$, for the rest of this section and also Appendix H, we use the training risk

$$\hat{R}_n(\beta, \mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \omega_i \log\Big(1 + e^{-y_i X_i^\intercal \beta}\Big) + \frac{\lambda}{2n}\|\beta\|^2.$$

D.4. Definitions

To complete the proof, we must first introduce the various terminology and techniques that are used throughout, from smoothing the labels and minimum function to the continuous Lindeberg interpolation.

D.4.1. Smoothing the Labels

First we will define the way in which we smooth our labels and subsequently the risk function. To do so, let us define the mollifier $\zeta_\gamma : \mathbb{R} \to \mathbb{R}$ for $\gamma \in (0, 1)$ as

$$\zeta_\gamma(x) := \mathsf{C} \cdot \exp\left(\frac{\gamma^2}{x^2 - \gamma^2}\right) \cdot \mathbb{I}(|x| < \gamma),$$

where C is chosen such that $\int_{\mathbb{R}} \zeta_\gamma(x)\,\mathrm{d}x = 1$. Then for a given function $f : \mathbb{R} \to \mathbb{R}$ we define

$$f_\gamma := f * \zeta_\gamma$$

as the convolution of f with ζ_γ, noting that this makes f smooth. We can then define a smoothed version of the labels as

$$\eta_i := \mathbf{1}^{\pm}_{\gamma}\left(a_{\mathcal{B}_i}^{\intercal} X_{\mathcal{B}_i} \beta^* - \varepsilon_i\right). \tag{14}$$

From here we can define the new smoothed risk as

$$\hat{R}_n^\gamma(\beta; \mathbf{X}) := \frac{1}{n}\sum_{i=1}^{n} \omega_i \log\left(1 + e^{-\eta_i X_i^\intercal \beta}\right) + \frac{\lambda}{2n}\|\beta\|_2^2,$$

where we have replaced each y_i with its smoothed counterpart η_i.

D.4.2. Smoothing the Minimum

Next we will define the function that will be used to approximate the minimum over our parameter space. For the set $\tilde{\mathcal{S}}$, define the smoothed minimum

$$f_\delta : \mathbb{R}^{>0} \times \mathbb{R}^{n \times p} \to \mathbb{R}, \quad f_\delta(\alpha, \mathbf{X}) := \frac{-1}{n\alpha} \log\left(\sum_{\beta \in \tilde{\mathcal{S}}_\delta} \exp\left[-n\alpha \hat{R}_n^\gamma(\beta; \mathbf{X})\right]\right),$$

where the sum is over the minimal $\delta\sqrt{p}$-net $\tilde{\mathcal{S}}_\delta$. When α is fixed or understood from context, we will refer to $f_\delta(\alpha, \mathbf{X})$ as simply $f_\delta(\mathbf{X})$.

D.4.3. INTERPOLATION TECHNIQUE

Finally, we define the interpolation that we will use for the proof of our main theorem. For $t \in [0, \frac{\pi}{2}]$, let

$$\mathbf{U}^t := \sin(t)\mathbf{X} + \cos(t)\mathbf{G}.$$

When t is fixed or understood from context, we will refer to $\mathbf{U}^t$ as simply $\mathbf{U}$. Now, for each $i = 1, \dots, n$, define the weight functions

$$w_\gamma(\beta) := \frac{e^{-n\alpha\hat{R}_n^\gamma(\beta,\mathbf{U})}}{\sum_{\beta'\in\tilde{S}_\delta} e^{-n\alpha\hat{R}_n^\gamma(\beta',\mathbf{U})}}, \qquad w_\gamma^{i,k}(\beta) := \frac{e^{-n\alpha\hat{R}_n^{\gamma,i,k}(\beta,\mathbf{U})}}{\sum_{\beta'\in\tilde{S}_\delta} e^{-n\alpha\hat{R}_n^{\gamma,i,k}(\beta',\mathbf{U})}}.$$

Also define expectation with respect to the density induced by these weights as

$$\langle \mathsf{g}(\beta)\rangle := \sum_{\beta\in\tilde{S}_\delta} w_\gamma(\beta)\mathsf{g}(\beta), \qquad \langle \mathsf{g}(\beta)\rangle_{i,k} := \sum_{\beta\in\tilde{S}_\delta} w_\gamma^{i,k}(\beta)\mathsf{g}(\beta), \tag{15}$$

where

$$\hat{R}_n^{\gamma,i,k}(\beta;\boldsymbol{\eta},\mathbf{U}) := \frac{1}{n}\sum_{j\notin\mathcal{B}_i} \omega_j \log\left(1 + e^{-\eta_j U_j^\top \beta}\right) + \frac{\lambda}{2n}\|\beta\|_2^2$$

represents the risk taken only over the points outside the block $\mathcal{B}_i$ containing X_i. Also for each $i = 1, \dots, n$ define the conditional expectation

$$\mathbb{E}_{(i,k)}[\,\cdot\,] := \mathbb{E}[\,\cdot\, \mid \mathbf{U}^{ik}],$$

where $\mathbf{U}^{ik}$ is used to denote the interpolation matrix without $\mathcal{B}_i$:

$$\mathbf{U}^{ik} := (U_1, \dots, U_{i-1}, 0, \dots, 0, U_{i+k}, \dots, U_n),$$

noting that this forces $\mathbf{U}^{ik} \perp\!\!\!\perp U_{\mathcal{B}_i}$. Lastly, we also define a "gradient" term

$$\mathcal{D}_i(U_{\mathcal{B}_i},\beta) := \left(\eta_i\omega_i\sigma_{i\beta}\right)\beta + \left(\sum_{j\in\mathcal{B}_i} \omega_j\sigma_{j\beta}\eta_j' a_{ji} U_j^\top\beta\right)\beta^* \in \mathbb{R}^p,$$

where $\sigma_{i\beta} := \sigma(-\eta_i U_i{}^\top\beta)$ and $\eta_i' := \mathbf{1}_\gamma^{\pm\prime}(a_{\mathcal{B}_i}^\top U_{\mathcal{B}_i}\beta^* - \varepsilon_i)$. When the data matrix is clear from context, we will write $\mathcal{D}_i(U_{\mathcal{B}_i},\beta)$ as simply $\mathcal{D}_i(\beta)$. Lastly, we use the shorthand

$$\ell(a,b) := \log\left(1 + e^{-ab}\right), \qquad \ell_i(\beta) := \ell(\eta_i, U_i^\top\beta).$$

With these, we are ready to begin the proof.

D.5. Proof of the Theorem 15

In this subsection we finally prove Theorem 15 which, from our previous remarks, immediately proves (5) of Theorem 2.

Proof of Theorem 15 For ease, let us refer to the quantity of interest as

$$d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}_n(\beta;\mathbf{G})\right) = (\star).$$

Let $\alpha, \delta, \gamma, \tau > 0$. We may first bound

$$\begin{aligned}(\star) &\leq d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{X})\right) + d_{\mathcal{H}}\left(\min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{G})\right) \qquad (16)\\ &+ d_{\mathcal{H}}\left(\min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{G}), \min_{\beta} \hat{R}_n(\beta;\mathbf{G})\right)\\ &\overset{(i)}{\leq} 2\mathsf{C}_1\sqrt{k\gamma} + d_{\mathcal{H}}\left(\min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{G})\right), \qquad (17)\end{aligned}$$

where (i) follows from applying Lemma 22 to the first and third summands. Then, we use Lemma 23 to bound

$$d_{\mathcal{H}}\left(\min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}^{\gamma}_n(\beta;\mathbf{G})\right) \leq d_{\mathcal{H}}\left(f_{\delta}(\alpha,\mathbf{X}), f_{\delta}(\alpha,\mathbf{G})\right) + \mathsf{C}_2\left(\sqrt{k}\delta + \frac{1}{\alpha}\log\left(\frac{1}{\delta}\right)\right). \qquad (18)$$

From here, we will prove universality for $f_{\delta}(\mathbf{X})$, and then show why this is sufficient. Recall from above the interpolator

$$\mathbf{U}^t := \sin(t)\mathbf{X} + \cos(t)\mathbf{G}, \quad t \in [0, \tfrac{\pi}{2}].$$

By the fundamental theorem of calculus, since $\mathbf{U}^0 = \mathbf{G}$ and $\mathbf{U}^{\pi/2} = \mathbf{X}$, we may bound

$$\left|\mathbb{E}\left[h\left(f_{\delta}(\mathbf{X})\right) - h\left(f_{\delta}(\mathbf{G})\right)\right]\right| \leq \int_0^{\pi/2} \left|\mathbb{E}\left[\partial_t h\left(f_{\delta}(\mathbf{U}^t)\right)\right]\right| \mathrm{d}t.$$

Using the chain rule we may expand

$$\partial_t h\left(f_{\delta}(\mathbf{U})\right) = \frac{-h'\left(f_{\delta}(\mathbf{U})\right)}{n}\sum_{i=1}^{n}\langle \tilde{U}_i^{\intercal}\mathcal{D}_i(\beta)\rangle$$

where we set

$$\tilde{\mathbf{U}}^t := \partial_t \mathbf{U}^t = \cos(t)\mathbf{X} - \sin(t)\mathbf{G}.$$

Using Lemma 24 we obtain that

$$\lim_{n\to\infty}\left|\mathbb{E}\left[h\left(f_{\delta}(\mathbf{X})\right) - h\left(f_{\delta}(\mathbf{G})\right)\right]\right| \overset{(i)}{\leq} \int_0^{\pi/2} \limsup_{n\to\infty}\left|\mathbb{E}\left[\partial_t h\left(f_{\delta}(\mathbf{U}^t)\right)\right]\right| \mathrm{d}t \overset{(ii)}{\leq} \frac{\pi}{2}\tau, \qquad (19)$$

where (i) is from the Dominated Convergence Theorem with the dominating function given by the bound on this derivative in Theorem 26, and (ii) is from Lemma 24. Combining (19) with (17) and (18), we conclude that

$$\lim_{n\to\infty}(\star) \leq \mathsf{C}\left[\sqrt{k\gamma} + \sqrt{k}\delta + \frac{1}{\alpha}\log\left(\frac{1}{\delta}\right) + \tau\right].$$

Noting that the left-hand side is independent of our four parameters $(\alpha, \delta, \gamma, \tau)$, we can take our limits in the proper order to conclude that

$$\lim_{n\to\infty} (\star) \leq \lim_{\delta\to 0} \lim_{\substack{\tau,\gamma\to 0\\ \alpha\to\infty}} \mathsf{C}\left[\sqrt{k\gamma} + \sqrt{k}\delta + \frac{1}{\alpha}\log\left(\frac{1}{\delta}\right) + \tau\right] = 0.$$

■

Appendix E. Proof of Theorem 3 (7) under Assumption 8(i): Training risk universality under m-dependence

The next result restates Theorem 3 (7) under Assumption 8(i), i.e. the universality universality of the training risk in the m-dependent setting:

Theorem 16 (Training risk universality under m-dependence) *Let $(X_i, y_i(X_i))_{i=1}^n$ and $(G_i, y_i(G_i))_{i=1}^n$ be generated under Assumptions 2-4, 8(i), and 9, where each $G_i \sim \mathcal{N}(\mathbf{0}, \mathrm{Var}(X_i))$. Then*

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_n(\beta; \mathbf{G})\right) \to 0. \tag{20}$$

Proof of Theorem 16 Let $M \in \mathbb{Z}^+$ be fixed. Define new matrices $\mathbf{X}^M, \mathbf{G}^M \in \mathbb{R}^{n' \times p}$ as

$$\begin{aligned} \mathbf{X}^M &:= (X_1, \ldots, X_M, X_{M+m+1}, \ldots, X_{2M+m}, X_{2M+2m+1}, \ldots)^\intercal \\ \mathbf{G}^M &:= (G_1, \ldots, G_M, G_{M+m+1}, \ldots, G_{2M+m}, G_{2M+2m+1}, \ldots)^\intercal, \end{aligned}$$

noting that

$$n' \in [n - (r+1)m + 1,\ n - rm] \subset \left[n\frac{M}{M+m} - m,\ n\frac{M}{M+m} + m\right] = [nq - m, nq + m], \tag{21}$$

where $r := \lfloor \frac{n}{M+m} \rfloor$ and $q := \frac{M}{M+m}$. By construction, $\mathbf{X}^M$ and $\mathbf{G}^M$ are block dependent with block size M. We may also define

$$\begin{aligned} \mathbf{X}^m &:= (X_{M+1}, \ldots, X_{M+m}, X_{2M+m+1}, \ldots, X_{2M+2m}, \ldots) \\ \mathbf{G}^m &:= (G_{M+1}, \ldots, G_{M+m}, G_{2M+m+1}, \ldots, G_{2M+2m}, \ldots) \end{aligned}$$

so that every vector X_i is either in $\mathbf{X}^M$ or $\mathbf{X}^m$. For simplicity we can also write these indexing sets as

$$\begin{aligned} B_M &:= \{1, \ldots, M, M+m+1, \ldots, 2M+m, \ldots, \} \\ B_m &:= [n] \setminus B_M. \end{aligned}$$

If we once again refer to our quantity of interest as

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_n(\beta; \mathbf{G})\right) = (\star),$$

then we may bound

$$(\star) \leq \underbrace{d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta; \mathbf{X}), \min_{\beta} \hat{R}_{n'}(\beta; \mathbf{X}^M)\right)}_{(a)} + \underbrace{d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_{n'}(\beta; \mathbf{X}^M), \min_{\beta} \hat{R}_{n'}(\beta; \mathbf{G}^M)\right)}_{(b)}$$
$$+ \underbrace{d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_{n'}(\beta; \mathbf{G}^M), \min_{\beta} \hat{R}_n(\beta; \mathbf{G})\right)}_{(c)}.$$

By Theorem 2, we know that $(b) \to 0$ as $n \to \infty$. Hence we have

$$\limsup_{n\to\infty} (\star) \leq \limsup_{n\to\infty} (a) + \limsup_{n\to\infty} (c).$$

We may now apply Theorem 29 with $\tilde{m} = m$ to say that for some $\mathsf{C}_d > 0$,

$$\limsup_{n\to\infty} (\star) \leq 2\mathsf{C}_d \frac{m}{M} \sqrt{M+m} = O\left(M^{-1/2}\right). \tag{22}$$

As this holds for every $M \in \mathbb{Z}^+$, we take $M \to \infty$ to obtain the result. ■

Appendix F. Proof of Theorem 3 (7) under Assumption 8(*ii*): Training risk universality under mixing

We first prove the universality of the train risk if the data is made of blocks of size k that are almost independent. We will then combine this result with Theorem 29 to get the desired result.

Assumption 13 (β_{mix} almost independent blocks of size k) *For all i, j such that $j \geq i+k$ we have that if we define $\mathcal{A} := \sigma((X_\ell, y_\ell)_{\ell \leq i})$ and $\mathcal{B} := \sigma((X_\ell, y_\ell)_{\ell > j})$ then $\beta\big(\mathcal{A}, \mathcal{B}\big) \leq \beta_{\text{mix}}$, where $\beta\big(\mathcal{A}, \mathcal{B}\big)$ is the β−mixing coefficient between the sigma-algebras $\mathcal{A}$ and $\mathcal{B}$ (see Bradley (2005) for a definition).*

Under this assumption we can establish the universality of the train risk.

Lemma 17 *Assume that $(\mathbf{X}_i, y_i)$ and $(\mathbf{G}_i, y_i(\mathbf{G}_i))$ satisfy Assumption 2-5. Assume that* $\mathrm{Var}((\mathbf{G}_\mathrm{i})) = \mathrm{Var}((\mathbf{X}_\mathrm{i}))$. *Moreover assume that Assumption 13 also holds. Then we have that there is a constant C such that*

$$\limsup_{n\to\infty} d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n(\beta; \mathbf{G})\right) \leq C\beta_{\text{mix}}^{1/15}$$

Proof For ease, let us refer to the quantity of interest as

$$d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n(\beta; \mathbf{G})\right) = (\star).$$

Let $\alpha, \delta, \gamma, \tau > 0$. We may first bound

$$\begin{aligned}(\star) &\leq d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{X})\right) + d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{G})\right) \\ &\quad + d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{G}), \min_{\beta} \hat{R}_n(\beta; \mathbf{G})\right) \\ &\overset{(i)}{\leq} 2\mathsf{C}_1 \sqrt{\gamma} \frac{\sqrt{\max(\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2, \mathbb{E}\|\mathbf{G}\|_{\mathcal{S}_p}^2)}}{\sqrt{n}} + d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{G})\right), \end{aligned} \tag{23}$$

where (i) follows from applying Lemma 22 to the first and third summands. Then, we use Lemma 23 to bound

$$d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta; \mathbf{G})\right) \tag{24}$$

$$\leq d_{\mathcal{H}}\left(f_\delta(\alpha, \mathbf{X}), f_\delta(\alpha, \mathbf{G})\right) + \mathsf{C}_2 \left(\frac{\sqrt{\max(\mathbb{E}\|\mathbf{X}\|_{\mathcal{B}(0,\sqrt{p})}^2, \mathbb{E}\|\mathbf{G}\|_{\mathcal{B}(0,\sqrt{p})}^2)}}{\sqrt{n}} \delta + \frac{1}{\alpha} \log\left(\frac{1}{\delta}\right) \right). \tag{25}$$

Using the bound in Corollary 20 under Assumption 8(ii) we know that there exists a constant $\mathsf{C}_3 > 0$ such that

$$\frac{\sqrt{\max(\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2, \mathbb{E}\|\mathbf{G}\|_{\mathcal{S}_p}^2)}}{\sqrt{n}}, \frac{\sqrt{\max(\mathbb{E}\|\mathbf{X}\|_{\mathcal{B}(0,\sqrt{p})}^2, \mathbb{E}\|\mathbf{G}\|_{\mathcal{B}(0,\sqrt{p})}^2)}}{\sqrt{n}} \leq \mathsf{C}_3 \mathcal{S}.$$

Hence we have

$$(\star) \leq d_{\mathcal{H}}\left(f_\delta(\alpha, \mathbf{X}), f_\delta(\alpha, \mathbf{G})\right) + \mathsf{C}_2 \left(\mathsf{C}_3 \mathcal{S} \delta + \frac{1}{\alpha} \log\left(\frac{1}{\delta}\right) \right) + 2\mathsf{C}_1 \mathsf{C}_3 \sqrt{\gamma} \mathcal{S}. \tag{26}$$

From here, we will prove universality for $f_\delta(\mathbf{X})$, and then show why this is sufficient. Recall from above the interpolator

$$\mathbf{U}^t := \sin(t)\mathbf{X} + \cos(t)\mathbf{G}, \quad t \in [0, \tfrac{\pi}{2}].$$

By the fundamental theorem of calculus, since $\mathbf{U}^0 = \mathbf{G}$ and $\mathbf{U}^{\pi/2} = \mathbf{X}$, we may bound

$$\left|\mathbb{E}\left[h\left(f_\delta(\mathbf{X})\right) - h\left(f_\delta(\mathbf{G})\right)\right]\right| \leq \int_0^{\pi/2} \left|\mathbb{E}\left[\partial_t h\left(f_\delta(\mathbf{U}^t)\right)\right]\right| \mathrm{d}t.$$

Using the chain rule we may expand

$$\partial_t h\left(f_\delta(\mathbf{U})\right) = \frac{-h'\left(f_\delta(\mathbf{U})\right)}{n} \sum_{i=1}^n \langle \tilde{U}_i^\intercal \mathcal{D}_i(\beta) \rangle$$

To further bound this, we will replace specific blocks ($X_{\mathcal{B}_i}$) with independent blocks, allowing us to use the results established in Appendix H.3. More precisely, we will show that for an average $i \leq n$ the expectation $\mathbb{E}\Big[h'\left(f_\delta(\mathbf{U})\right) \langle \tilde{U}_i^\intercal \mathcal{D}_i(\beta) \rangle\Big]$ is approximately the same if the block containing the i-th observation $\left(X_j, y_j(X_j)\right)_{j \in \mathcal{B}_i}$ is independent from the others. We also use the notation $j > \mathcal{B}_i$ to denote the set of indices after the last index of the block $\mathcal{B}_i$, and analogously use $j < \mathcal{B}_i$ for those before $\mathcal{B}_i$.

In this goal, for $i \leq n$ we write (W^i_j) the process such that $\left(W^i_j, y_j(W^i_j)\right)_{j \in \mathcal{B}_i}$, $\left(W^i_j, y_j(W^i_j)\right)_{j > \mathcal{B}_i}$ and $\left(W^i_j, y_j(W^i_j)\right)_{j < \mathcal{B}_i}$ are independent and have the same marginals as $\left(X_j, y_j(X_j)\right)_{j \in \mathcal{B}_i}$, $\left(X_j, y_j(X_j)\right)_{j > \mathcal{B}_i}$ and $\left(X_j, y_j(X_j)\right)_{j < \mathcal{B}_i}$. We define similarly $\left(G^i_j, y_j(G^i_j)\right)_{j \in \mathcal{B}_i}$, $\left(G^i_j, y_j(G^i_j)\right)_{j > \mathcal{B}_i}$ and $\left(G^i_j, y_j(G^i_j)\right)_{j < \mathcal{B}_i}$. The interpolated process between W^i and $\mathbf{G}^i$ is written as

$$U^{(i)}_j := \mathrm{sin(t)W^i_j + cos(t)G^i_j}, \qquad \mathrm{\tilde{U}^{(i)}_j := cos(t)W^i_j - sin(t)G^i_j}.$$

Denote $\mathcal{D}^{(i)}_j(\beta)$ the version of $\mathcal{D}_i(\beta)$ for $\mathbf{U}^{(i)}$ and $\tilde{\mathbf{U}}^{(i)}$. Similarly we write $\langle \cdot \rangle_{(i)}$ the version of $\langle \cdot \rangle$ for $U^{(i)}$ (see (15) for the definition of $\langle \cdot \rangle$).

Choose $L > 0$ to be a real. Define

$$\mathcal{D}_{i,L}(\beta) := \mathcal{D}_i(\beta)\ \mathbb{I}(|\tilde{U}_i^T \beta|, |U_i^T \beta| \leq L)$$

and

$$\mathcal{D}^{(i)}_{i,L}(\beta) := \mathcal{D}^{(i)}_i(\beta) \mathbb{I}(|\tilde{(}U^{(i)}_i)^T \beta|, |(U^{(i)}_i)^T \beta| \leq L).$$

We first remark that we can switch $\mathcal{D}_i(\beta)$ for the truncated $\mathcal{D}_{i,L}(\beta)$ without changing too much the value of the expectation $\mathbb{E}\Big(-h'\,(f_\delta(\mathbf{U}))\langle \tilde{U}_i^\intercal \mathcal{D}_i(\beta)\rangle\Big)$. Indeed we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \Big|\mathbb{E}\Big(-h'\,(f_\delta(\mathbf{U}))\langle \tilde{U}_i^\intercal \mathcal{D}_{i,L}(\beta)\rangle\Big) - \mathbb{E}\Big(-h'\,(f_\delta(\mathbf{U}))\langle \tilde{U}_i^\intercal \mathcal{D}_i(\beta)\rangle\Big)\Big| \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(\langle |\tilde{U}_i^\intercal \mathcal{D}_i(\beta)| \mathbb{I}(|\tilde{U}_i^T\beta| > L)\rangle\Big) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(\langle |\tilde{U}_i^\intercal \mathcal{D}_i(\beta)| \mathbb{I}(|U_i^T\beta| > L)\rangle\Big) \\
&\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(\langle \Big\{|\tilde{U}_i^\intercal\beta| + |U_i^\intercal\beta||\tilde{U}_i^\intercal\beta^*|\Big\}\ \mathbb{I}(|\tilde{U}_i^T\beta| > L)\rangle\Big) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big(\langle \Big\{|\tilde{U}_i^\intercal\beta| + |U_i^\intercal\beta||\tilde{U}_i^\intercal\beta^*|\Big\}\ \mathbb{I}(|U_i^T\beta| > L)\rangle\Big) \\
&\overset{(i)}{\leq} \frac{1}{nL^{1/4}} \sum_{i=1}^n \mathbb{E}\Big(\langle \Big\{|\tilde{U}_i^\intercal\beta|^{5/4} + |U_i^\intercal\beta||\tilde{U}_i^\intercal\beta|^{1/4}|\tilde{U}_i^\intercal\beta^*|\Big\}\rangle\Big) \\
&\qquad + \frac{1}{nL^{1/4}} \sum_{i=1}^n \mathbb{E}\Big(\langle \Big\{|\tilde{U}_i^\intercal\beta||U_i^T\beta|^{1/4} + |U_i^\intercal\beta|^{5/4}|\tilde{U}_i^\intercal\beta^*|\Big\}\rangle\Big) \\
&\overset{(ii)}{\leq} \frac{1}{nL^{1/4}} \sum_{i=1}^n \mathbb{E}\Big(\langle |\tilde{U}_i^\intercal\beta|^{5/4}\rangle\Big) + L^{-1/4} \max_{i\leq n} \|\tilde{U}_i^\intercal\beta\|_6 \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\langle (U_i^T\beta)^2\rangle)} \Big\{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\langle (U_i^T\beta)^{3/4}\rangle)\Big\}^{3/4} \\
&\qquad + \frac{1}{L^{1/4}} \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\langle (\tilde{U}_i^T\beta)^2\rangle)} \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\langle |\tilde{U}_i^T\beta|^{3/4}\rangle^{2/3})} + \max_i \|\tilde{U}_i^T\beta^*\|_3 \sqrt{\frac{1}{n}\sum_{i=1}^n \mathbb{E}(\langle (U_i^T\beta)^2\rangle)} \\
&\leq \frac{1}{L^{1/4}} \Big\{\mathbb{E}\Big(n^{-1}\|\tilde{U}\|^2_{\mathcal{S}_p}\Big)^{5/8} + \max_{i\leq n} \|\tilde{U}_i^\intercal\beta\|_6 \sqrt{\mathbb{E}\Big(n^{-1}\|\mathbf{U}\|^2_{\mathcal{S}_p}\Big)} \mathbb{E}\Big(n^{-1}\|\tilde{U}\|^2_{\mathcal{S}_p}\Big)^{9/32} \\
&\qquad + \sqrt{\mathbb{E}\Big(n^{-1}\|\tilde{U}\|^2_{\mathcal{S}_p}\Big)}\Big(\mathbb{E}\Big(n^{-1}\|U\|^2_{\mathcal{S}_p}\Big)\Big)^{1/8} + \max_i \|\tilde{U}_i^T\beta^*\|_3 \sqrt{\mathbb{E}\Big(n^{-1}\|U\|^2_{\mathcal{S}_p}\Big)}\Big\}.
\end{aligned}$$

where (i) is a consequence of the fact that $\mathbb{I}(|\tilde{U}_i^T\beta| \geq L) \leq \frac{|\tilde{U}_i^T\beta|^{1/4}}{L^{1/4}}$, (ii) comes from Jensen inequality combined with Hölder inequality. Hence using the bound in Corollary 20 under Assumption 8(ii) and using Assumption 3 we have that there exists a constant C_4 such that

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\Big(-h'\left(f_\delta(\mathbf{U})\right)\langle\tilde{U}_i^\intercal\mathcal{D}_{i,L}(\beta)\rangle\Big) - \mathbb{E}\Big(-h'\left(f_\delta(\mathbf{U})\right)\langle\tilde{U}_i^\intercal\mathcal{D}_i(\beta)\rangle\Big)\right| \leq \mathsf{C}_4 L^{-1/4}. \tag{27}$$

Moreover according to Lemma 38 we have that there is a constant C_5 such that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\Big(-h'\left(f_\delta(\mathbf{U})\right)\langle\tilde{U}_i^\intercal\mathcal{D}_{i,L}(\beta)\rangle\Big) - \mathbb{E}\Big(-h'\left(f_\delta(\mathbf{U}^{(i)})\right)\langle(\tilde{\mathbf{U}}^{(i)})^\intercal\mathcal{D}_{i,L}^{(i)}(\beta)\rangle_{(i)}\Big)\right| \qquad (28)\\
&\overset{(i)}{\leq} 4(3\beta_{\text{mix}})^{1/3}\frac{1}{n}\sum_{i=1}^{n}\Big\{L(1+\|\tilde{U}_i^T\beta^*\|_{3/2}\Big\}\\
&\overset{(ii)}{\leq} L\mathsf{C}_5(\beta_{\text{mix}})^{1/3}
\end{aligned}$$

where (i) is due to the fact that $|\tilde{U}_i\mathcal{D}_{i,L}(\beta)| \leq |\tilde{U}_i\beta|\ \mathbb{I}(|\tilde{U}_i\beta| \leq L) + \mathbb{I}(|U_i\beta| \leq L)|\ \tilde{U}_i\beta^*||U_i^T\beta|$ combined with the triangle inequality (ii) is due to assumption 3 which imply that $\limsup_n \sup_i \|U_i^T\beta^*\|_{3/2} < \infty$. Hence combining Theorem 24 with Eqs. (27) and (28) we obtain that

$$\frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\Big(\frac{-h'\left(f_\delta(\mathbf{U})\right)}{n}\sum_{i=1}^{n}\langle\tilde{U}_i^\intercal\mathcal{D}_i(\beta)\rangle\Big)\right| \leq L\mathsf{C}_5\beta_{\text{mix}}^{1/3} + \tau + \mathsf{C}_4 L^{-1/4}. \tag{29}$$

This directly implies that

$$\begin{aligned}
\lim_{n\to\infty}\left|\mathbb{E}\left[h\left(f_\delta(\mathbf{X})\right) - h\left(f_\delta(\mathbf{G})\right)\right]\right| &\overset{(i)}{\leq} \int_0^{\pi/2}\limsup_{n\to\infty}\left|\mathbb{E}\left[\partial_t h\left(f_\delta(\mathbf{U}^t)\right)\right]\right|\mathrm{d}t \qquad (30)\\
&\overset{(ii)}{\leq} \frac{\pi}{2}\Big[L\mathsf{C}_5\beta_{\text{mix}}^{1/3} + \tau + \mathsf{C}_4 L^{-1/4}\Big] \qquad (31)
\end{aligned}$$

where (i) is from the Dominated Convergence Theorem with the dominating function given by the bound on this derivative in Theorem 26, and (ii) is from Eq. (29). Combining (31) with (23) and (26), we conclude that

$$\lim_{n\to\infty}(\star) \leq \mathsf{C}\left[\sqrt{k\gamma} + \sqrt{k}\delta + \frac{1}{\alpha}\log\left(\frac{1}{\delta}\right) + \tau + L\beta_{\text{mix}}^{1/3} + L^{-1/4}\right].$$

Noting that the left-hand side is independent of our four parameters $(\alpha, \delta, \gamma, \tau)$, we can take our limits in the proper order to conclude that

$$\lim_{n\to\infty}(\star) \leq \lim_{\delta\to 0}\lim_{\substack{\tau,\gamma\to 0\\ \alpha\to\infty}}\mathsf{C}\left[\sqrt{k\gamma} + \sqrt{k}\delta + \frac{1}{\alpha}\log\left(\frac{1}{\delta}\right) + \tau\right] = C\left[L\beta_{\text{mix}}^{1/3} + L^{-1/4}\right].$$

Optimizing over $L \geq 0$ gives us the desired result.

■

The next result restates Theorem 3 (7) under Assumption 8(ii), i.e. training risk universality under mixing:

Theorem 18 (Universality of the training risk under Assumption 8(ii)) *Let $(X_i, y_i(X_i))_{i=1}^n$ and $(G_i, y_i(G_i))_{i=1}^n$ be generated under Assumptions 2-4, 8(ii), and 9, where each $G_i \sim \mathcal{N}(\mathbf{0}, \mathrm{Var}(X_i))$. Then*

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_n(\beta; \mathbf{G})\right) \to 0. \tag{32}$$

Proof of Theorem 18 Let $M, \tilde{m} \in \mathbb{Z}^+$ be fixed. Define new matrices $\mathbf{X}^M, \mathbf{G}^M \in \mathbb{R}^{n' \times p}$ as

$$\begin{aligned}
\mathbf{X}^M &:= (X_1, \ldots, X_M, X_{M+\tilde{m}+1}, \ldots, X_{2M+\tilde{m}}, X_{2M+2\tilde{m}+1}, \ldots)^\top \\
\mathbf{G}^M &:= (G_1, \ldots, G_M, G_{M+\tilde{m}+1}, \ldots, G_{2M+\tilde{m}}, G_{2M+2\tilde{m}+1}, \ldots)^\top,
\end{aligned}$$

noting that

$$n' \in [n - (r+1)\tilde{m} + 1,\ n - r\tilde{m}] \subset \left[n\frac{M}{M+\tilde{m}} - \tilde{m},\ n\frac{M}{M+\tilde{m}} + \tilde{m}\right] = [nq - \tilde{m}, nq + \tilde{m}], \tag{33}$$

where $r := \lfloor \frac{n}{M+\tilde{m}} \rfloor$ and $q := \frac{M}{M+\tilde{m}}$. By construction, $\mathbf{X}^M$ and $\mathbf{G}^M$ satisfy Assumption 13 with $\beta_{\mathrm{mix}} = \beta(\tilde{m})$. We may also define

$$\begin{aligned}
\mathbf{X}^{\tilde{m}} &:= (X_{M+1}, \ldots, X_{M+\tilde{m}}, X_{2M+\tilde{m}+1}, \ldots, X_{2M+2\tilde{m}}, \ldots) \\
\mathbf{G}^{\tilde{m}} &:= (G_{M+1}, \ldots, G_{M+\tilde{m}}, G_{2M+\tilde{m}+1}, \ldots, G_{2M+2\tilde{m}}, \ldots)
\end{aligned}$$

so that every vector X_i is either in $\mathbf{X}^M$ or $\mathbf{X}^{\tilde{m}}$. For simplicity we can also write these indexing sets as

$$\begin{aligned}
B_M &:= \{1, \ldots, M, M + \tilde{m} + 1, \ldots, 2M + \tilde{m}, \ldots, \} \\
B_{\tilde{m}} &:= [n] \setminus B_M.
\end{aligned}$$

If we once again refer to our quantity of interest as

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_n(\beta; \mathbf{G})\right) = (\star),$$

then we may bound

$$\begin{aligned}
(\star) \le\ & \underbrace{d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right)}_{(a)} + \underbrace{d_{\mathcal{H}}\left(\min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M), \min_\beta \hat{R}_{n'}(\beta; \mathbf{G}^M)\right)}_{(b)} \\
& + \underbrace{d_{\mathcal{H}}\left(\min_\beta \hat{R}_{n'}(\beta; \mathbf{G}^M), \min_\beta \hat{R}_n(\beta; \mathbf{G})\right)}_{(c)}.
\end{aligned}$$

By Theorem 17, we know that there is $\mathsf{C} > 0$ such that $\limsup\ (b) \le \mathsf{C}\beta(\tilde{m})^{1/15}$ as $n \to \infty$. Hence we have

$$\limsup_{n\to\infty}\ (\star) \le \limsup_{n\to\infty}\ (a) + \limsup_{n\to\infty}\ (c) + \mathsf{C}\beta(\tilde{m})^{1/15}.$$

We may now apply Theorem 29 with $\tilde{m}$ to say that for some $\mathsf{C}_d > 0$,

$$\limsup_{n\to\infty} (\star) \leq 2\mathsf{C}_d \frac{m}{M} \sqrt{M+m} + \mathsf{C}\beta(\tilde{m})^{1/15} = O\Big(M^{-1/2} + \mathsf{C}\beta(\tilde{m})^{1/15}\Big). \tag{34}$$

As this holds for every $M \in \mathbb{Z}^+ +$, we take $M \to \infty$ to obtain that

$$\limsup_{n\to\infty} (\star) \leq \mathsf{C}\beta(\tilde{m})^{1/15}. \tag{35}$$

Finally noting that this holds for an arbitrary $\tilde{m}$ gives us the desired result. ■

Appendix G. Proof of Theorem 2 (6) and Theorem 3 (8): Test risk universality

In this section, we prove the second equation of both Theorems 2 and 3 concerning test risk universality, as both share the same proof once training risk universality is proved. We focus on presenting the proof for the 0-1 loss, i.e. the test risk R_{test} is defined with

$$\ell_{\text{test}}(X_{\text{new}}^\intercal \hat{\beta}\,,\, X_{\text{new}}^\intercal \beta^*) \;:=\; \mathbb{I}\Big\{\, \mathbb{I}\Big\{\sigma(X_{\text{new}}^\intercal \hat{\beta}) \geq \frac{1}{2}\Big\} = \mathbb{I}\{X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} \geq 0\}\,\Big\}$$

for both $\hat{\beta} = \hat{\beta}(\mathbf{X})$ and $\hat{\beta} = \hat{\beta}(\mathbf{G})$. As our proof strategy relies on approximating ℓ_{test} by the 1-Lipschitz functions in $\tilde{\mathcal{F}}$, the same proof also works if ℓ_{test} is already Lipschitz. Therefore the result also applies to any locally Lipschitz ℓ_{test}, which is Lipschitz over the compact set $\mathcal{S}_p$.

Proof of Theorem 2 (6) and Theorem 3 (8) Recall that $\sigma(x) = \frac{1}{1+e^{-x}}$. Our test loss can then be re-expressed as

$$\begin{aligned} R_{\text{test}}(\hat{\beta}) &= \mathbb{E}\Big[\, \mathbb{I}\Big\{\mathbb{I}\{\sigma(X_{\text{new}}^\intercal \hat{\beta}) \geq \tfrac{1}{2}\} = \mathbb{I}\{X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} \geq 0\}\Big\} \Big| \hat{\beta}\Big] \\ &= \mathbb{E}\Big[\, \mathbb{I}\Big\{\mathbb{I}\{X_{\text{new}}^\intercal \hat{\beta} \geq 0\} = \mathbb{I}\{X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} \geq 0\}\Big\} \Big| \hat{\beta}\Big] \\ &= \mathbb{E}\Big[\, \mathbb{I}\Big\{X_{\text{new}}^\intercal \hat{\beta} \geq 0\,,\, X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} \geq 0\Big\} + \mathbb{I}\Big\{X_{\text{new}}^\intercal \hat{\beta} < 0\,,\, X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} < 0\Big\} \Big| \hat{\beta}\Big]\,, \end{aligned} \tag{36}$$

and by a similar argument,

$$R_{\text{test}}^G(\hat{\beta}) \;=\; \mathbb{E}\Big[\, \mathbb{I}\Big\{G_{\text{new}}^\intercal \hat{\beta} \geq 0\,,\, G_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} \geq 0\Big\} + \mathbb{I}\Big\{G_{\text{new}}^\intercal \hat{\beta} < 0\,,\, G_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}} < 0\Big\} \Big| \hat{\beta}\Big]\,.$$

For convenience, we denote the random $\mathbb{R}^2$ vectors

$$V_X \;:= (X_{\text{new}}^\intercal \hat{\beta}\,,\, X_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}})^\intercal \qquad \text{and} \qquad V_G \;:= (G_{\text{new}}^\intercal \hat{\beta}\,,\, G_{\text{new}}^\intercal \beta^* - \varepsilon_{\text{new}})^\intercal\,.$$

We first perform a standard smoothing of the indicator function. By Lemma 34 of Huang et al. (2023), for any $\tau \in \mathbb{R}$ and $\delta > 0$, there exists a continuously differentiable function $h_{\tau;\delta}$ such that $h_{\tau+\delta;\delta}(x) \leq \mathbb{I}\{x \geq \tau\} \leq h_{\tau;\delta}(x)$ for all $x \in \mathbb{R}$ and that $\partial h_{\tau;\delta}$ is bounded in norm by δ^{-1}. Moreover $h_{\tau;\delta}$ takes value in $[0,1]$. We use this to construct the $\mathbb{R}^2 \to \mathbb{R}$ function $\tilde{h}_{\tau;\delta}(x,y) := h_{\tau;\delta}(x)\, h_{\tau;\delta}(y)$, which satisfies

$$\tilde{h}_{\delta;\delta}(x,y) \;\leq\; \mathbb{I}\{x \geq 0, y \geq 0\} \;=\; \mathbb{I}\{x \geq 0\}\,\mathbb{I}\{y \geq 0\} \;\leq\; \tilde{h}_{0;\delta}(x,y)\,.$$

This implies that for every $\delta > 0$, almost surely

$$\begin{aligned}
&\mathbb{E}\Big[\, \mathbb{I}\Big\{X_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \,-\, \mathbb{I}\Big\{G_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \Big|\, \hat{\beta}\Big] \\
&\leq\; \mathbb{E}[\tilde{h}_{0;\delta}(V_X) - \tilde{h}_{\delta;\delta}(V_G) \,|\, \hat{\beta}] \\
&\leq\; \mathbb{E}[\tilde{h}_{0;\delta}(V_X) - \tilde{h}_{0;\delta}(V_G) + \tilde{h}_{0;\delta}(V_G) - \tilde{h}_{\delta;\delta}(V_G) \,|\, \hat{\beta}] \\
&\leq\; \mathbb{E}[\tilde{h}_{0;\delta}(V_X) - \tilde{h}_{0;\delta}(V_G) + \mathbb{I}\{(V_G)_1 \geq -\delta\,,\, (V_G)_2 \geq -\delta\} - \mathbb{I}\{(V_G)_1 \geq \delta\,,\, (V_G)_2 \geq \delta\} \,|\, \hat{\beta}] \\
&\leq\; \mathbb{E}[\tilde{h}_{0;\delta}(V_X) - \tilde{h}_{0;\delta}(V_G) \,|\, \hat{\beta}] + \mathbb{P}((V_G)_1 \in [-\delta, \delta) \,|\, \hat{\beta}) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,|\, \hat{\beta})\;.
\end{aligned}$$

In the last inequality, we have noted that if $(V_G)_1, (V_G)_2 \geq -\delta$ is true and yet $(V_G)_1, (V_G)_2 \geq \delta$ is false, we must have either $(V_G)_1 \in [-\delta, \delta)$ or $(V_G)_2 \in [-\delta, \delta)$. Now let $\delta \in (0, 1]$. Notice that $\delta\, \tilde{h}_{0;\delta} \in \tilde{\mathcal{F}}$, where $\tilde{\mathcal{F}}$ is defined in Assumption 6. Also note that $\hat{\beta}$ is independent of X_{new} and G_{new} in V_X and V_G. This implies according to Assumption 6 that

$$\begin{aligned}
&\Big|\,\mathbb{E}[\tilde{h}_{0;\delta}(V_X) \,|\, \hat{\beta}] - \mathbb{E}[\tilde{h}_{0;\delta}(V_G) \,|\, \hat{\beta}]\,\Big| \\
&\leq \frac{1}{\delta}\; \sup_{f \in \tilde{\mathcal{F}}} \Big|\mathbb{E}\Big[f(X_{\text{new}}^{\intercal}\hat{\beta}, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}}) \,\Big|\, \hat{\beta}\Big] - \mathbb{E}\Big[f(G_{\text{new}}^{\intercal}\hat{\beta}, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}}) \,\Big|\, \hat{\beta}\Big]\Big| \\
&\leq \frac{1}{\delta}\; \sup_{f \in \tilde{\mathcal{F}}}\; \sup_{\beta \in \mathcal{S}_p} \Big|\mathbb{E}\Big[f(X_{\text{new}}^{\intercal}\hat{\beta}, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}})\Big] - \mathbb{E}\Big[f(G_{\text{new}}^{\intercal}\hat{\beta}, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}})\Big]\Big| \\
&\overset{(a)}{\leq} \frac{1}{\delta}\; \sup_{f \in \tilde{\mathcal{F}}}\; \sup_{\beta \in \mathcal{S}_p} \Big|\mathbb{E}\Big[f(X_{\text{new}}^{\intercal}\hat{\beta}, X_{\text{new}}^{\intercal}\beta^{*})\Big] - \mathbb{E}\Big[f(G_{\text{new}}^{\intercal}\hat{\beta}, G_{\text{new}}^{\intercal}\beta^{*})\Big]\Big| \;\;=:\;\; \frac{1}{\delta}\Delta_n
\end{aligned}$$

In (a), we have used a conditioning on ε_{new}, moved the suprema and the norm inside the expectation over ε_{new} and observed that the function $f(\,\bullet\,, \bullet - \varepsilon_{\text{new}}) \in \tilde{\mathcal{F}}$ almost surely. Substituting this into the above yields that, almost surely

$$\begin{aligned}
&\mathbb{E}\Big[\, \mathbb{I}\Big\{X_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \,-\, \mathbb{I}\Big\{G_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \Big|\, \hat{\beta}\Big] \\
&\leq\; \frac{1}{\delta}\Delta_n + \mathbb{P}((V_G)_1 \in [-\delta, \delta) \,|\, \hat{\beta}) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,|\, \hat{\beta})\;.
\end{aligned}$$

By a similar argument, we can obtain that almost surely

$$\begin{aligned}
&\mathbb{E}\Big[\, \mathbb{I}\Big\{G_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \,-\, \mathbb{I}\Big\{X_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \Big|\, \hat{\beta}\Big] \\
&\leq\; \mathbb{E}[\tilde{h}_{0;\delta}(V_G) - \tilde{h}_{\delta;\delta}(V_X) \,|\, \hat{\beta}] \\
&\leq\; \mathbb{E}[\tilde{h}_{\delta;\delta}(V_G) - \tilde{h}_{\delta;\delta}(V_X) + \tilde{h}_{0;\delta}(V_G) - \tilde{h}_{\delta;\delta}(V_G) \,|\, \hat{\beta}] \\
&\leq\; \mathbb{E}[\tilde{h}_{\delta;\delta}(V_X) - \tilde{h}_{\delta;\delta}(V_G) \,|\, \hat{\beta}] + \mathbb{P}((V_G)_1 \in [-\delta, \delta) \,|\, \hat{\beta}) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,|\, \hat{\beta}) \\
&\leq\; \frac{1}{\delta}\Delta_n + \mathbb{P}((V_G)_1 \in [-\delta, \delta) \,|\, \hat{\beta}) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,|\, \hat{\beta})\;.
\end{aligned}$$

Combining the two bounds implies that, almost surely,

$$\begin{aligned}
(\star) \;\;:=\;\; &\Big|\,\mathbb{E}\Big[\, \mathbb{I}\Big\{X_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, X_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \,-\, \mathbb{I}\Big\{G_{\text{new}}^{\intercal}\hat{\beta} \geq 0\,,\, G_{\text{new}}^{\intercal}\beta^{*} - \varepsilon_{\text{new}} \geq 0\Big\} \Big|\, \hat{\beta}\Big]\Big| \\
&\leq\; \frac{2}{\delta}\Delta_n + \mathbb{P}((V_G)_1 \in [-\delta, \delta) \,|\, \hat{\beta}) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,|\, \hat{\beta})\;.
\end{aligned}$$

To control the probability terms, notice that conditioning on $\hat{\beta}$, $(V_G)_1 = G_{\text{new}}^{\intercal}\hat{\beta} \,|\, \hat{\beta} \sim \mathcal{N}(0, \hat{\beta}^{\intercal}\Sigma_{\text{new}}\hat{\beta})$ and $(V_G)_2 \,|\, \varepsilon_{\text{new}}, \hat{\beta} \;\sim\; \mathcal{N}(-\varepsilon_{\text{new}}, \beta^{*\intercal}\Sigma_{\text{new}}\beta^{*})$. Therefore by a standard anti-concentration result for

Gaussians (see e.g. Carbery and Wright (2001)), there is an absolute constant $C'' > 0$ such that, almost surely,

$$\mathbb{P}((V_G)_1 \in [-\delta, \delta) \,\big|\, \hat\beta) + \mathbb{P}((V_G)_2 \in [-\delta, \delta) \,\big|\, \beta\,,\, \varepsilon_{\text{new}}) \;\leq\; C''\delta\Big(\frac{1}{\hat\beta^\intercal \Sigma_{\text{new}}\hat\beta} + \frac{1}{\beta^{*\intercal}\Sigma_{\text{new}}\beta^*}\Big)\,.$$

Meanwhile by Assumption 7, for every $\epsilon > 0$,

$$\mathbb{P}(D_\epsilon(\mathbf{G}) > 0) \;\to\; 1\,, \qquad \text{where} \quad D_\epsilon(\mathbf{G}) \;:=\; \min_{\beta \in \mathcal{S}_p\,,\,|(\beta^\intercal\Sigma_{\text{new}}\beta)^{1/2} - \bar\chi| > \epsilon} \hat R_n(\beta; \mathbf{G}) - \min_{\beta\in\mathcal{S}_p} \hat R_n(\beta;\mathbf{G})\,,$$

and by the universality of the training risk (Theorem 2 (5) or Theorem 3 (7)), we also have $\mathbb{P}(D_\epsilon(\mathbf{X}) > 0) \to 1$. This implies that

$$\begin{aligned}\mathbb{P}(\,\big|(\hat\beta(\mathbf{X})^\intercal\Sigma_{\text{new}}\hat\beta(\mathbf{X}))^{1/2} - \bar\chi\big| \leq \epsilon\,) \;&= \mathbb{P}(D_\epsilon(\mathbf{X}) > 0) \;\to\; 1\,,\\ \mathbb{P}(\,\big|(\hat\beta(\mathbf{G})^\intercal\Sigma_{\text{new}}\hat\beta(\mathbf{G}))^{1/2} - \bar\chi\big| \leq \epsilon\,) \;&= \mathbb{P}(D_\epsilon(\mathbf{G}) > 0) \;\to\; 1\,.\end{aligned} \tag{37}$$

In other words, for both $\hat\beta = \hat\beta(\mathbf{X})$ and $\hat\beta = \hat\beta(\mathbf{G})$, $\hat\beta^\intercal\Sigma_{\text{new}}\hat\beta \xrightarrow{\mathbb{P}} \bar\chi^2$ in probability. Moreover, $\bar\chi > 0$ and $\beta^{*\intercal}\Sigma_{\text{new}}\beta^* \xrightarrow{\mathbb{P}} \chi_*^2 > 0$ by Assumption 7. This allows us to consider a rare event

$$E_\chi \;:=\; \Big\{\hat\beta^\intercal\Sigma_{\text{new}}\hat\beta < \frac{\bar\chi^2}{2}\,,\, \beta^{*\intercal}\Sigma_{\text{new}}\beta^* < \frac{\chi_*^2}{2}\Big\} \qquad \text{such that} \quad \mathbb{P}(E_\chi) \;\to\; 0\,.$$

Denoting E_χ^c as the complement of E_χ, we obtain that for any $\epsilon' > 0$,

$$\begin{aligned}\mathbb{P}(|(\star)| > \epsilon') \;&\leq \mathbb{P}\Big(\frac{2}{\delta}\Delta_n + C''\delta\Big(\frac{1}{\hat\beta^\intercal\Sigma_{\text{new}}\hat\beta} + \frac{1}{\beta^{*\intercal}\Sigma_{\text{new}}\beta^*}\Big) > \epsilon'\Big)\\ &\leq \mathbb{P}\Big(\frac{2}{\delta}\Delta_n + C''\delta\Big(\frac{1}{\hat\beta^\intercal\Sigma_{\text{new}}\hat\beta} + \frac{1}{\beta^{*\intercal}\Sigma_{\text{new}}\beta^*}\Big) > \epsilon' \,\Big|\, E_\chi^c\Big) + \mathbb{P}(E_\chi)\\ &\leq \mathbb{I}\Big\{\frac{2}{\delta}\Delta_n + C''\delta\Big(\frac{2}{\bar\chi^2} + \frac{2}{\chi_*^2}\Big) > \epsilon'\Big\} + \mathbb{P}(E_\chi)\,.\end{aligned}$$

By Assumption 6, $\Delta_n \to 0$. Since the above is valid for any δ, whose choice is independent of ϵ', we can choose $\delta = \sqrt{\Delta_n}$, which implies that the above converge to zero. In other words, we have shown that

$$\Big|\mathbb{E}\Big[\,\mathbb{I}\Big\{X_{\text{new}}^\intercal\hat\beta \geq 0\,,\, X_{\text{new}}^\intercal\beta^* - \varepsilon_{\text{new}} \geq 0\Big\} \,-\, \mathbb{I}\Big\{G_{\text{new}}^\intercal\hat\beta \geq 0\,,\, G_{\text{new}}^\intercal\beta^* - \varepsilon_{\text{new}} \geq 0\Big\}\,\Big|\,\hat\beta\Big]\Big| \;\xrightarrow{\mathbb{P}}\; 0$$

for both $\hat\beta = \hat\beta(\mathbf{X})$ and $\hat\beta = \hat\beta(\mathbf{G})$. By an exactly analogous argument, we have

$$\Big|\mathbb{E}\Big[\,\mathbb{I}\Big\{X_{\text{new}}^\intercal\hat\beta < 0\,,\, X_{\text{new}}^\intercal\beta^* - \varepsilon_{\text{new}} < 0\Big\}\,\Big] - \mathbb{E}\Big[\,\mathbb{I}\Big\{G_{\text{new}}^\intercal\hat\beta < 0\,,\, G_{\text{new}}^\intercal\beta^* - \varepsilon_{\text{new}} < 0\Big\}\,\Big|\,\hat\beta\Big]\Big| \;\to\; 0\,.$$

In view of (36), we can use a triangle inequality to obtain that

$$|R_{\text{test}}(\hat\beta(\mathbf{X})) - R_{\text{test}}^G(\beta(\mathbf{X}))| \;\xrightarrow{\mathbb{P}} 0 \qquad\qquad \text{and} \qquad\qquad |R_{\text{test}}(\hat\beta(\mathbf{G})) - R_{\text{test}}^G(\beta(\mathbf{G}))| \;\xrightarrow{\mathbb{P}} 0\,.$$

Meanwhile, note that $R_{\text{test}}^G(\hat\beta)$ depends on $\hat\beta$ only through the mean-zero conditionally Gaussian variable $G_{\text{new}}^\intercal\hat\beta$, which is completely characterized by $\text{Var}[G_{\text{new}}^\intercal\hat\beta\,|\,\hat\beta] = \hat\beta^\intercal\Sigma_{\text{new}}\hat\beta$. In view of (37), both $\hat\beta(\mathbf{X})^\intercal\Sigma_{\text{new}}\hat\beta(\mathbf{X})$ and $\hat\beta(\mathbf{G})^\intercal\Sigma_{\text{new}}\hat\beta(\mathbf{G})$ converge in probaility to the same constant $\bar\chi^2$. This implies

$$|R_{\text{test}}^G(\hat\beta(\mathbf{X})) - R_{\text{test}}^G(\beta(\mathbf{G}))| \;\xrightarrow{\mathbb{P}} 0\,,$$

which in particular implies the desired statement that $|R_{\text{test}}(\hat\beta(\mathbf{X})) - R_{\text{test}}^G(\beta(\mathbf{G}))| \;\xrightarrow{\mathbb{P}}\; 0$ ■

Appendix H. Important Lemmas

In this section, we present the statements and proofs of the various lemmas used to prove our main theorems.

H.1. Auxiliary Lemmas

The lemma and its corollary aim to bound the expectation of the maximum possible norm of our signal $\mathbf{X}\beta$, conditional on the fact that a Bernstein-like Inequality holds.

Lemma 19 (Operator Norm Bound) *Let $(\mathbf{Y}_i)$ be a sequence of $\mathbb{R}^p$-valued random vectors, and let $\mathsf{R} > 0$. Suppose that $\frac{n}{p} \to \kappa$ and that there exist constants $\mathsf{K}, \mathsf{C}_1, \mathsf{c}_2, \mathsf{C}_3$ such that*

1. $\sup_{i \le n} \|\mathrm{Var}(\mathbf{Y}_i)\|_{\mathrm{op}} \le \frac{1}{p}\mathsf{K}$.
2. *For all $\beta \in \mathcal{S} := \mathcal{B}_p(0, \mathsf{R}\sqrt{p})$ and $t > 0$, we have*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}(\mathbf{Y}_i^\mathsf{T}\beta)^2 - \mathbb{E}((\mathbf{Y}_i^\mathsf{T}\beta)^2)\right| \ge t\right) \le \mathsf{C}_1 \exp\left(-\mathsf{c}_2 n\left(\frac{t}{\mathsf{C}_3\mathsf{R}^2} \wedge \frac{t^2}{\mathsf{C}_3^2\mathsf{R}^4}\right)\right). \tag{38}$$

Then there exists $\mathsf{C}_\mathsf{R} > 0$ depending on R and the constants above such that for n sufficiently large,

$$\mathbb{E}\left[\|\mathbf{Y}\|_{\mathcal{S}}^2\right] \le \mathsf{C}_\mathsf{R}\, p,$$

where $\|\mathbf{Y}\|_{\mathcal{S}}$ is as defined in (12).

Proof Let $\beta \in \mathcal{S}$, meaning by definition $\|\beta\|_2 \le \mathsf{R}\sqrt{p}$. Note that by the first statement of the lemma, we have that

$$\mathbb{E}\left[\|\mathbf{Y}\beta\|^2\right] = n\beta^\mathsf{T}\overline{\Sigma}_n\beta \le n\mathsf{R}^2\mathsf{K},$$

where $\overline{\Sigma}_n := \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(\mathbf{Y}_i)$. Now, we can note that for $s > 0$,

$$\mathbb{P}\left(\|\mathbf{Y}\beta\|^2 \ge s\right) = \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \ge \frac{s - n\beta^\mathsf{T}\overline{\Sigma}_n\beta}{n}\right),$$

where $W_i := (\mathbf{Y}_i^\mathsf{T}\beta)^2 - \beta^\mathsf{T}\mathrm{Var}(\mathbf{Y}_i)\beta$. If we let $\mathsf{M}_\mathsf{R} := \mathsf{R}^2(\mathsf{K} + \mathsf{C}_3)$, then by (38), we obtain that for $s > 0$,

$$\mathbb{P}\left(\|\mathbf{Y}\beta\|^2 \ge n\mathsf{M}_\mathsf{R}(s+1)\right) \le \mathsf{C}_1 \cdot \exp\left[-\mathsf{c}_2 n\left(s \wedge s^2\right)\right], \tag{39}$$

which follows from noting that

$$\frac{n\mathsf{M}_\mathsf{R}(s+1) - n\beta^\mathsf{T}\overline{\Sigma}_n\beta}{n\mathsf{C}_3\mathsf{R}^2} \ge \frac{\mathsf{M}_\mathsf{R}(s+1) - \mathsf{M}_\mathsf{R}}{\mathsf{M}_\mathsf{R}} = s.$$

Now let $\varepsilon > 0$, and define $\mathcal{S}_\varepsilon$ to be a minimal $\varepsilon\sqrt{p}$-net of $\mathcal{S}$. Also, for $t > 0$, define the quantity

$$\eta_t := \frac{\mathsf{C}_3\sqrt{p} + t}{\sqrt{n}} \quad \text{for} \quad \mathsf{C}_3 := \sqrt{\frac{1}{\mathsf{c}_2}\log\left(1 + \frac{3\mathsf{R}}{\varepsilon}\right)}.$$

Then if we set $s = \eta_t \vee \eta_t^2$ in (39), by a union bound we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{\beta \in \mathcal{S}_\varepsilon} \|\mathbf{Y}\beta\|^2 \geq n\mathsf{M}_\mathsf{R}\left((\eta_t \vee \eta_t^2) + 1\right)\right) &\overset{(i)}{\leq} \mathsf{C}_1 |\mathcal{S}_\varepsilon| \exp\left[-\mathsf{c}_2 n \cdot \eta_t^2\right] \\ &\overset{(ii)}{\leq} \mathsf{C}_1 \left(\frac{3\mathsf{R}}{\varepsilon}\right)^p \exp\left[-\mathsf{c}_2(\mathsf{C}_3^2 p + t^2)\right] \\ &\overset{(iii)}{\leq} \mathsf{C}_1 e^{-\mathsf{c}_2 t^2},\end{aligned}$$

where (i) is via the fact that for any $x \geq 0$, we have

$$(x \vee x^2) \wedge (x \vee x^2)^2 = x^2,$$

(ii) is via Corollary 4.2.13 of Vershynin (2018) bounding the cardinality of a minimal ε-net, and (iii) is from the definition of η_t and C_3. Now we may bound the error between the supremum on the whole space and the supremum on the ε-net by applying a similar technique to Lemma 4.4.1 of Vershynin (2018), which gives

$$\sup_{\beta \in \mathcal{S}} \|\mathbf{Y}\beta\|^2 \leq \frac{1}{1 - 2\varepsilon} \sup_{\beta \in \mathcal{S}_\varepsilon} \|\mathbf{Y}\beta\|^2.$$

We conclude that, for $\mathsf{A} := \sqrt{\mathsf{M}_\mathsf{R}/(1 - 2\varepsilon)}$,

$$\mathbb{P}\left(\|\mathbf{Y}\|_\mathcal{S}^2 \geq n\mathsf{A}^2\left((\eta_t \vee \eta_t^2) + 1\right)\right) \leq \mathsf{C}_1 e^{-\mathsf{c}_2 t^2}. \tag{40}$$

Now, let us define the two events

$$\mathcal{E}_1 := \left\{\frac{\|\mathbf{Y}\|_\mathcal{S}^2}{n\mathsf{A}^2} - 1 \leq \eta_t \vee \eta_t^2\right\}, \qquad \mathcal{E}_2 := \left\{\frac{\|\mathbf{Y}\|_\mathcal{S}}{\sqrt{n}\mathsf{A}} \leq 1\right\},$$

where we note that $\mathcal{E}_1$ is exactly the high-probability event of (40). Then we can first see that, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, we have

$$\mathcal{E}_1 \cap \mathcal{E}_2 \implies \|\mathbf{Y}\|_\mathcal{S} \leq \mathsf{A}\sqrt{n}, \tag{41}$$

which is simply from the definition of the event $\mathcal{E}_2$. On the other hand, for the event $\mathcal{E}_1 \cap \mathcal{E}_2^c$, we have

$$\begin{aligned}\left(\frac{\|\mathbf{Y}\|_\mathcal{S}}{\sqrt{n}\mathsf{A}} - 1\right)^2 \vee \left|\frac{\|\mathbf{Y}\|_\mathcal{S}}{\sqrt{n}\mathsf{A}} - 1\right| &\overset{(i)}{\leq} \left|\frac{\|\mathbf{Y}\|_\mathcal{S}^2}{n\mathsf{A}^2} - 1\right| \\ &\overset{(ii)}{=} \frac{\|\mathbf{Y}\|_\mathcal{S}^2}{n\mathsf{A}^2} - 1 \\ &\overset{(iii)}{\leq} \eta_t \vee \eta_t^2,\end{aligned}$$

where (i) is from the fact that $(x - y)^2 \vee |x - y| \leq |x^2 - y^2|$ for $x, y > 0$ and $x + y \geq 1$, (ii) follows from $\mathcal{E}_2^c$, and (iii) from $\mathcal{E}_1$. Since we also know that

$$(x \vee x^2) \leq (y \vee y^2) \implies x \leq y \quad \text{for} \quad x, y \geq 0,$$

we can say that

$$\mathcal{E}_1 \cap \mathcal{E}_2^c \implies \left| \frac{\|\mathbf{Y}\|_{\mathcal{S}}}{\sqrt{n}\mathsf{A}} - 1 \right| \leq \eta_t \implies \|\mathbf{Y}\|_{\mathcal{S}} \leq \mathsf{A}\left(\sqrt{n} + \mathsf{C}_3\sqrt{p} + t\right). \tag{42}$$

Combining (41) and (42), we conclude that

$$\mathcal{E}_1 = (\mathcal{E}_1 \cap \mathcal{E}_2) \cup (\mathcal{E}_1 \cap \mathcal{E}_2^c) \implies \|\mathbf{Y}\|_{\mathcal{S}_p} \leq \tilde{\mathsf{A}}\left(\sqrt{n} + \sqrt{p} + t\right),$$

where $\tilde{\mathsf{A}} = \mathsf{A}(\mathsf{C}_3 + 1)$. Thus we have

$$\begin{aligned} \mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}} \geq \tilde{\mathsf{A}}\left(\sqrt{n} + \sqrt{p} + t\right)\right) &\leq \mathbb{P}\left(\mathcal{E}_1^c\right) \\ &\leq \mathsf{C}_1 e^{-\mathsf{c}_2 t^2}. \end{aligned}$$

If we set the variable $y = \tilde{\mathsf{A}}^2\left(\sqrt{n} + \sqrt{p} + t\right)^2$, then we have that

$$\mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}} \geq \sqrt{y}\right) \leq \mathsf{C}_1 \cdot \exp\left[-\mathsf{c}_2\left(\frac{\sqrt{y}}{\tilde{\mathsf{A}}} - \sqrt{n} - \sqrt{p}\right)^2\right]$$

whenever $\sqrt{y} \geq \tilde{\mathsf{A}}\left(\sqrt{n} + \sqrt{p}\right)$. This lets us conclude via the tail-integral formula that, after setting $\varepsilon = 1/4$ for example,

$$\begin{aligned} \mathbb{E}\|\mathbf{Y}\|_{\mathcal{S}}^2 &= \int_0^\infty \mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}}^2 \geq y\right) \mathrm{d}y \\ &= \int_0^\infty \mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}} \geq \sqrt{y}\right) \mathrm{d}y \\ &= \int_0^{\tilde{\mathsf{A}}^2(\sqrt{n}+\sqrt{p})^2} \mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}} \geq \sqrt{y}\right) \mathrm{d}y + \int_{\tilde{\mathsf{A}}^2(\sqrt{n}+\sqrt{p})^2}^\infty \mathbb{P}\left(\|\mathbf{Y}\|_{\mathcal{S}} \geq \sqrt{y}\right) \mathrm{d}y \\ &\leq \int_0^{\tilde{\mathsf{A}}^2(\sqrt{n}+\sqrt{p})^2} 1 \,\mathrm{d}y + \mathsf{C}_1 \int_{\tilde{\mathsf{A}}^2(\sqrt{n}+\sqrt{p})^2}^\infty \exp\left[-\mathsf{c}_2\left(\frac{\sqrt{y}}{\tilde{\mathsf{A}}} - \sqrt{n} - \sqrt{p}\right)^2\right] \mathrm{d}y \\ &\leq \tilde{\mathsf{A}}^2(\sqrt{n} + \sqrt{p})^2 + \tilde{\mathsf{A}}^2\left(\frac{1}{\mathsf{c}_2} + \frac{\sqrt{\pi}(\sqrt{n} + \sqrt{p})}{\sqrt{\mathsf{c}_2}}\right) \\ &\leq \mathsf{C}_{\mathsf{R}} \cdot p, \end{aligned}$$

for

$$\mathsf{C}_{\mathsf{R}} \propto \mathsf{R}^2 \log(1 + 12\mathsf{R}).$$

■

Now that we have given a bound on this scaled operator norm, we can use this to obtain a bound in the specific cases of block dependence, m-dependence, and mixing.

Corollary 20 (Dependent Norm Bound) *Let* $(\mathbf{X}_i)$ *be a sequence of* $\mathbb{R}^p$*-valued random vectors satisfying Assumptions 2-4. Then setting* $\mathcal{S}$ *in Lemma 19 as either* $\mathcal{B}_p(0, \sqrt{p})$ *or* $\mathcal{S}_p$ *as in (4), the following holds for n sufficiently large:*

(1) *If* $(\mathbf{X}_i)$ *satisfies Assumption 1 (block-dependence), then* $\mathbb{E}[\|\mathbf{X}\|_{\mathcal{S}}^2] \leq \mathsf{C}kp$.

(2) *If* $(\mathbf{X}_i)$ *satisfies Assumption 8(i) (m-dependence), then* $\mathbb{E}[\|\mathbf{X}\|_{\mathcal{S}}^2] \leq \mathsf{C}mp$.

(3) *If* $(\mathbf{X}_i)$ *satisfies Assumption 8(ii) (β-mixing) and there exists* $\varepsilon > 0$ *such that*

$$\mathsf{S} := \sum_{\ell=1}^{\infty} \beta(\ell)^{\frac{\varepsilon}{2+\varepsilon}} < \infty,$$

then $\mathbb{E}[\|\mathbf{X}\|_{\mathcal{S}}^2] \leq \mathsf{CS}p$. *Furthermore, for* $(\mathbf{G}_i)$ *also satisfying Assumptions 2-4 with* $\mathbf{G}_i \sim \mathcal{N}(0, \mathrm{Var}(\mathbf{X}_i))$ *and* $\mathrm{cov}(\mathbf{G}) = \mathrm{cov}(\mathbf{X})$, *we have* $\mathbb{E}[\|\mathbf{G}\|_{\mathcal{S}}^2] \leq \mathsf{CS}p$.

Proof We display the proofs for $\mathcal{S}_p$, as those for the ball of radius $\sqrt{p}$ are identical.

(1): Note that (1) follows from (2), as block-dependence with block parameter k is a special case of m-dependence with $m = k$.

(2): We check that the two conditions of Lemma 19 hold. For the first one, note that by the sub-Gaussian Assumption 3 and Lemma 44, we have

$$\|\mathrm{Var}(\mathbf{X}_i)\|_{\mathrm{op}} \leq \frac{\mathsf{C}_1\mathsf{K}_X^2}{n} \leq \frac{\mathsf{B}}{p}$$

for some $\mathsf{C}_1, \mathsf{B} > 0$ and n sufficiently large. For the second condition, by sub-Gaussianity and the definition of $\mathcal{S}_p$, we have that

$$\|(\mathbf{X}_i^\top\beta)^2\|_{\psi_1} \overset{(i)}{=} \|\mathbf{X}_i^\top\beta\|_{\psi_2}^2 \leq \frac{\mathsf{K}_X^2\|\beta\|^2}{n} \leq \mathsf{C}_2\mathsf{L}^2 \tag{43}$$

for n sufficiently large, where (i) is by Lemma 2.7.6 of Vershynin (2018). Thus we may apply Lemma 36 with $Z_i := (X_i^\top\beta^2) - \mathbb{E}\left[(X_i^\top\beta^2)\right]$ and $\mathsf{K} = \mathsf{C}_2\mathsf{L}^2$. This gives us our second condition, namely (38), with the constants $(\mathsf{R}, \mathsf{K}, \mathsf{C}_1, \mathsf{c}_2, \mathsf{C}_3)$ set as $(\mathsf{L}, \mathsf{B}, 4, cm^{-1}, \mathsf{C}_2)$. For the Gaussin case, the first condition holds for the exact same reason. For the second condition, we again use Lemma 39, which allows us to satisfy (38) with the constants $(\mathsf{R}, \mathsf{K}, \mathsf{C}_1, \mathsf{c}_2, \mathsf{C}_3)$ set as $(\mathsf{L}, \mathsf{B}, 2, c\mathsf{S}^{-1}, \mathsf{K}_X\underline{c}^{-1})$, respectively.

(4): As above, the first condition of Lemma 19 holds for the same reason as (2) and (3). The second condition holds by the second statement of Lemma 39, which allows us to satisfy (38) with the constants $(\mathsf{R}, \mathsf{K}, \mathsf{C}_1, \mathsf{c}_2, \mathsf{C}_3)$ set as $(\mathsf{L}, \mathsf{B}, 2, c\mathsf{S}^{-1}, c_2\left(\tilde{C}_2 \wedge \tilde{C}_2'\right))$. ■

Next, recall that we have smoothed our labels $y_1, \ldots, y_n$ by taking a convolution of the sign function with a mollifier ζ_γ as in (14). This next lemma shows that the derivative of this convolution grows at a rate inversely proportional to the smoothing factor, γ.

Lemma 21 (Smoothed Label Derivative Bound) *Define* $\eta_i' := \mathbf{1}_\gamma^{\pm\prime}(a_{\mathcal{B}_i}^\top X_{\mathcal{B}_i}\beta^* - \varepsilon_i)$. *Then*

$$\left|\eta_i'\right| \leq 3\gamma^{-1}.$$

Proof Recall by the properties of convolution that a derivative can be "absorbed" into the convolution like so:

$$\begin{aligned}\eta_i' := \mathbf{1}_\gamma^{\pm\prime}(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^* - \varepsilon_i) &= (\mathbf{1}^\pm * \zeta_\gamma)'(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^* - \varepsilon_i) \\ &= (\mathbf{1}^\pm * \zeta_\gamma')(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^* - \varepsilon_i) \\ &= \int_{-\gamma}^{\gamma} \mathbf{1}^\pm(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^* - \varepsilon_i - t)\zeta_\gamma'(t)\,\mathrm{d}t.\end{aligned}$$

We may thus bound

$$\begin{aligned}\left|\eta_i'\right| &= \left|\int_{-\gamma}^{\gamma} \mathbf{1}^\pm(a_{\mathcal{B}_i}^\intercal X_{\mathcal{B}_i}\beta^* - \varepsilon_i - t)\zeta_\gamma'(t)\,\mathrm{d}t\right| \\ &\le \int_{-\gamma}^{\gamma} \left|\zeta_\gamma'(t)\right|\mathrm{d}t \\ &= 2\mathsf{C}\gamma^2 \int_{-\gamma}^{\gamma} \frac{|t|}{(t^2-\gamma^2)^2}\exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right)\mathrm{d}t \\ &\overset{(i)}{\le} 4\mathsf{C}\gamma^2 \int_0^{\gamma} \frac{t}{(t^2-\gamma^2)^2}\exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right)\mathrm{d}t \\ &= \frac{2}{e}\mathsf{C},\end{aligned}$$

where (*i*) is from the fact that the integrand is even. Thus it suffices to bound C, which is the integrating constant of our mollifier ζ_γ. We may lower bound the integral like so:

$$\begin{aligned}\mathsf{C}^{-1} &= \int_{-\gamma}^{\gamma} \exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right)\mathrm{d}t \\ &= 2\int_0^{\gamma} \exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right)\mathrm{d}t \\ &\overset{(i)}{\ge} 2\int_0^{\gamma/2} \exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right)\mathrm{d}t \\ &\overset{(ii)}{\ge} 2\int_0^{\gamma/2} \exp\left(\frac{\gamma^2}{(\gamma/2)^2-\gamma^2}\right)\mathrm{d}t \\ &= \gamma \cdot e^{-4/3} \\ &\ge \frac{\gamma}{4},\end{aligned}$$

where (*i*) follows from the fact that the integrand is non-negative, and (*ii*) from the fact that it is a decreasing function, as it has derivative

$$\frac{-2t\gamma^2}{(t^2-\gamma^2)^2}\exp\left(\frac{\gamma^2}{t^2-\gamma^2}\right) < 0 \quad \text{for} \quad 0 < t < \gamma.$$

We conclude that

$$\left|\eta_i'\right| \le \frac{2}{e}\cdot\frac{4}{\gamma} \le \frac{3}{\gamma}.$$

■

H.2. General lemmas

H.2.1. REPLACING THE TRUE MINIMUM WITH A SMOOTHED & DISCRETIZED MINIMUM

To prove that the two minimum risks are close in distribution to one another, we must first smooth the labels, and then also discretize the parameter space that we are taking the minimum over. The following two lemmas show that the error incurred by these two approximations is negligible in the limit.

Lemma 22 (Smoothing the Risk) *Suppose that* $\mathbf{X}$ *and* $\mathbf{G}$ *satisfy Assumptions 2–4. Let* $\gamma \in (0,1)$. *Then there exists* $\mathsf{C} > 0$ *such that for n sufficiently large,*

$$d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n(\beta;\mathbf{X}), \min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n^{\gamma}(\beta;\mathbf{X})\right) \le \mathsf{C}\frac{\sqrt{\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2}}{\sqrt{n}}\sqrt{\gamma}$$

$$d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n(\beta;\mathbf{G}), \min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n^{\gamma}(\beta;\mathbf{G})\right) \le \mathsf{C}\frac{\sqrt{\mathbb{E}\|\mathbf{G}\|_{\mathcal{S}_p}^2}}{\sqrt{n}}\sqrt{\gamma}.$$

Proof We show the proof only for $\mathbf{X}$, and note that the exact same technique holds for $\mathbf{G}$. Since $h \in \mathcal{H}$ is Lipschitz, we know that if $\tilde{\beta}$ and $\tilde{\beta}_\gamma$ are the minimizers of $\hat{R}_n$ and $\hat{R}_n^\gamma$ on $\tilde{\mathcal{S}}$ respectively, then

$$\begin{aligned} d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n(\beta;\mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta;\mathbf{X})\right) &= \sup_{h\in\mathcal{H}} \mathbb{E}\left[h\left(\hat{R}_n(\tilde{\beta};\mathbf{X})\right) - h\left(\hat{R}_n^{\gamma}(\tilde{\beta}_\gamma;\mathbf{X})\right)\right] \\ &\le \sup_{h\in\mathcal{H}} \|h'\|_\infty \mathbb{E}\left|\hat{R}_n(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}_\gamma;\mathbf{X})\right| \\ &\le \mathbb{E}\left|\hat{R}_n(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}_\gamma;\mathbf{X})\right|, \end{aligned}$$

where the last line is by definition of $\mathcal{H}$. To control this term, we can note that

$$\left|\hat{R}_n(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}_\gamma;\mathbf{X})\right| \le \max\left\{\underbrace{\left|\hat{R}_n(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X})\right|}_{(a)}, \underbrace{\left|\hat{R}_n(\tilde{\beta}_\gamma;\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}_\gamma;\mathbf{X})\right|}_{(b)}\right\}.$$

We bound the first term (a) like so:

$$\begin{aligned} \left|\hat{R}_n(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X})\right| &\overset{(i)}{\le} \frac{1}{n}\sum_{i=1}^n \left|\log\left(1+e^{-y_i X_i^\mathsf{T}\tilde{\beta}}\right) - \log\left(1+e^{-\eta_i X_i^\mathsf{T}\tilde{\beta}}\right)\right| \\ &\overset{(ii)}{\le} \frac{1}{n}\sum_{i=1}^n \left|X_i^\mathsf{T}\tilde{\beta}\right||y_i - \eta_i| \\ &\overset{(iii)}{\le} \frac{1}{n}\|\mathbf{X}\tilde{\beta}\|\,\|\boldsymbol{y}-\boldsymbol{\eta}\| \\ &\overset{(iv)}{\le} \frac{1}{n}\|\mathbf{X}\|_{\mathcal{S}_p}\|\boldsymbol{y}-\boldsymbol{\eta}\|, \end{aligned}$$

where (i) uses that $0 \le \omega_i \le 1$ for all $i \le n$, (ii) comes from treating the loss as a function of the label and Taylor expanding, (iii) is from Cauchy-Schwarz, and (iv) is from the definition of $\|\mathbf{X}\|_{\mathcal{S}_p}$. Applying Cauchy-Schwarz once more we see that

$$\mathbb{E}[(a)] \le \frac{1}{n}\sqrt{\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2 \mathbb{E}\|\boldsymbol{y}-\boldsymbol{\eta}\|^2}.$$

From here we note that

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{y}-\boldsymbol{\eta}\|^2 &= \sum_{i=1}^{n}\mathbb{E}[(y_i-\eta_i)^2]\\
&\overset{(i)}{=} \sum_{i=1}^{n}\mathbb{E}[(y_i-\eta_i)^2\mathbf{1}_{|a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*-\varepsilon_i|\le\gamma}]\\
&\overset{(ii)}{\le} \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{P}\left(|a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*-\varepsilon_i|\le\gamma \mid X_{\mathcal{B}_i}\right)\right]\\
&\overset{(iii)}{\le} n\gamma,
\end{aligned}$$

where the indicator in (i) is introduced as

$$y_i-\eta_i\neq 0 \iff |a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*-\varepsilon_i|\le\gamma,$$

(ii) is because $(y_i-\eta_i)^2\le 1$, and (iii) from

$$\begin{aligned}
\mathbb{P}\left(|a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*-\varepsilon_i|\le\gamma \mid X_{\mathcal{B}_i}\right) &= \sigma(a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*+\gamma)-\sigma(a_{\mathcal{B}_i}^{\top}X_{\mathcal{B}_i}\beta^*-\gamma)\\
&\le 2\gamma\|\sigma'\|_\infty\\
&\le \gamma.
\end{aligned}$$

We conclude that

$$\mathbb{E}[(a)] \le \frac{\sqrt{\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2}}{n}\sqrt{n\gamma} \le \frac{\sqrt{\mathbb{E}\|\mathbf{X}\|_{\mathcal{S}_p}^2}}{\sqrt{n}}\sqrt{\gamma}$$

for n sufficiently large. To finish, note that this exact string of inequalities also holds for $\mathbb{E}[(b)]$. ■

Now that we have bounded the difference between the original risk and smoothed risk in terms of the smoothing parameter γ, we can show that universality for the smoothed risk reduces to universality for the smooth minimum f_δ.

Lemma 23 (Discretization & Smooth-Min) *Let $\alpha,\delta>0$. Suppose that $\mathbf{X}$ and $\mathbf{G}$ satisfy Assumptions 2-4. Write $\mathcal{S} := \mathcal{B}(0,\sqrt{p})$. Then there exists $\mathsf{C}>0$ such that for n sufficiently large,*

$$\begin{aligned}
&d_{\mathcal{H}}\left(\min_{\beta}\hat{R}_n^{\gamma}(\beta;\mathbf{X}),\min_{\beta}\hat{R}_n^{\gamma}(\beta;\mathbf{G})\right)\\
&\le d_{\mathcal{H}}\left(f_\delta(\alpha,\mathbf{X}),f_\delta(\alpha,\mathbf{G})\right)+\mathsf{C}\Big(\delta+\frac{\delta}{\sqrt{n}}\sqrt{\mathbb{E}\left[\|\boldsymbol{X}\|_{\mathcal{S}}^2\right]\vee\mathbb{E}\left[\|\boldsymbol{G}\|_{\mathcal{S}}^2\right]}+\frac{1}{\alpha}\log\left(\frac{1}{\delta}\right)\Big)
\end{aligned}$$

Proof By the Triangle Inequality, we know that

$$\begin{aligned}
d_{\mathcal{H}}\left(\min_{\beta} \hat{R}_n^{\gamma}(\beta;\mathbf{X}), \min_{\beta} \hat{R}_n^{\gamma}(\beta;\mathbf{G})\right) &\leq d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n^{\gamma}(\beta;\mathbf{X}), \min_{\beta\in\tilde{\mathcal{S}}_\delta} \hat{R}_n^{\gamma}(\beta;\mathbf{X})\right) \\
&\quad + d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}_\delta} \hat{R}_n^{\gamma}(\beta;\mathbf{X}), f_\delta(\alpha,\mathbf{X})\right) \\
&\quad + d_{\mathcal{H}}\left(f_\delta(\alpha,\mathbf{X}), f_\delta(\alpha,\mathbf{G})\right) \\
&\quad + d_{\mathcal{H}}\left(f_\delta(\alpha,\mathbf{G}), \min_{\beta\in\tilde{\mathcal{S}}_\delta} \hat{R}_n^{\gamma}(\beta;\mathbf{G})\right) \\
&\quad + d_{\mathcal{H}}\left(\min_{\beta\in\tilde{\mathcal{S}}_\delta} \hat{R}_n^{\gamma}(\beta;\mathbf{G}), \min_{\beta\in\tilde{\mathcal{S}}} \hat{R}_n^{\gamma}(\beta;\mathbf{G})\right) \\
&=: \mathbb{D}_1 + \mathbb{D}_2 + \mathbb{D}_3 + \mathbb{D}_4 + \mathbb{D}_5.
\end{aligned}$$

For $\mathbb{D}_1$, let $\tilde{\beta}$ be the minimizer of the risk on $\tilde{\mathcal{S}}$, $\tilde{\beta}_\delta$ the closest point to it on the $\delta\sqrt{p}$-net $\tilde{\mathcal{S}}_\delta$, and $\tilde{\beta}'$ the minimizer of the risk on $\tilde{\mathcal{S}}_\delta$. Then we have that

$$\begin{aligned}
\mathbb{D}_1 &:= \sup_{h\in\mathcal{H}} \left|\mathbb{E}\left[h\left(\hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X})\right) - h\left(\hat{R}_n^{\gamma}(\tilde{\beta}';\mathbf{X})\right)\right]\right| \\
&\leq \sup_{h\in\mathcal{H}} \mathbb{E}\left|h\left(\hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X})\right) - h\left(\hat{R}_n^{\gamma}(\tilde{\beta}';\mathbf{X})\right)\right| \\
&\overset{(i)}{\leq} \mathbb{E}\left|\hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}';\mathbf{X})\right| \\
&\overset{(ii)}{\leq} \mathbb{E}\left|\hat{R}_n^{\gamma}(\tilde{\beta};\mathbf{X}) - \hat{R}_n^{\gamma}(\tilde{\beta}_\delta;\mathbf{X})\right| \\
&\leq \mathbb{E}\left[\sup_{\nu} \left|\langle\nabla\hat{R}_n^{\gamma}(\nu;\mathbf{X}), \tilde{\beta} - \tilde{\beta}_\delta\rangle\right|\right],
\end{aligned}$$

where (i) is from the definition of $\mathcal{H}$ and (ii) is because

$$\hat{R}_n^{\gamma}(\tilde{\beta}_\delta) \geq \hat{R}_n^{\gamma}(\tilde{\beta}')$$

by definition. The gradient of our risk has coordinates

$$\frac{\partial}{\partial\nu_j}\hat{R}_n^{\gamma}(\nu;\mathbf{X}) = \frac{\lambda}{n}\nu_j - \frac{1}{n}\sum_{i=1}^{n} \omega_i\eta_i X_{ij}\sigma_{i\nu},$$

and thus

$$
\begin{aligned}
\mathbb{D}_1 &\leq \frac{\lambda}{n}\mathbb{E}\left[\sup_{v}\left|\langle v, \tilde{\beta} - \tilde{\beta}_\delta\rangle\right|\right] + \frac{1}{n}\mathbb{E}\left[\sup_{v}\left|\left\langle \sum_{i=1}^{n} \eta_i\omega_i X_i \sigma_{iv}, \tilde{\beta} - \tilde{\beta}_\delta\right\rangle\right| \mathbb{I}(\mathbf{X} \in E_n)\right] \\
&\overset{(i)}{\leq} \frac{\lambda \mathsf{L}\delta p}{n} + \frac{1}{n}\mathbb{E}\left[\sup_{v}\left|\sum_{i=1}^{n}\eta_i\omega_i\sigma_{iv}\left\langle X_i, \tilde{\beta} - \tilde{\beta}_\delta\right\rangle\right| \mathbb{I}(\mathbf{X} \in E_n)\right] \\
&\overset{(ii)}{\leq} \frac{\lambda \mathsf{L}\delta p}{n} + \frac{1}{n}\mathbb{E}\left[\sup_{v}\sqrt{\sum_{i=1}^{n}\eta_i^2\sigma_{iv}^2\omega_i^2}\sqrt{\sum_{i=1}^{n}\left\langle X_i, \tilde{\beta} - \tilde{\beta}_\delta\right\rangle^2}\right] \\
&\overset{(iii)}{\leq} \frac{\lambda \mathsf{L}\delta p}{n} + \frac{1}{\sqrt{n}}\mathbb{E}\left[\sqrt{\sum_{i=1}^{n}\left\langle X_i, \tilde{\beta} - \tilde{\beta}_\delta\right\rangle^2}\right] \\
&\overset{(iv)}{\leq} \frac{\lambda \mathsf{L}\delta p}{n} + \frac{\delta}{\sqrt{n}}\mathbb{E}\left[\|\mathbf{X}\|_{\mathcal{S}}\right]
\end{aligned}
$$

Above, (i) is from the fact that

$$
\left|\langle v, \tilde{\beta} - \tilde{\beta}_\delta\rangle\right| \leq \|v\|\,\|\tilde{\beta} - \tilde{\beta}_\delta\| \leq \mathsf{L}\sqrt{p}\delta\sqrt{p} = \mathsf{L}\delta p,
$$

(ii) is a result of Cauchy-Schwarz, (iii) uses that $\eta_i, \omega_i, \sigma_{i,v} \leq 1$ for all $i \leq n$, (iv) is via the definition of $\|\mathbf{X}\|_{\mathcal{S}}$ and the fact that $\delta^{-1}(\tilde{\beta} - \tilde{\beta}_\delta) \in \mathcal{S}$. To bound $\mathbb{D}_2$, we observe that

$$
\begin{aligned}
\left| f_\delta(\alpha, \mathbf{X}) - \min_{\beta \in \tilde{\mathcal{S}}_\delta} \hat{R}_n^\gamma(\beta; \mathbf{X}) \right| &= \left| \frac{-1}{n\alpha}\log\left(\sum_{\beta \in \tilde{\mathcal{S}}_\delta}\exp\left[-n\alpha\hat{R}_n^\gamma(\beta; \mathbf{X})\right]\right) - \hat{R}_n^\gamma(\tilde{\beta}; \mathbf{X})\right| \\
&= \frac{1}{n\alpha}\left|\log\left(\sum_{\beta \in \tilde{\mathcal{S}}_\delta}\exp\left[-n\alpha\hat{R}_n^\gamma(\beta; \mathbf{X})\right]\right) - \log\left(\exp\left[-n\alpha\hat{R}_n^\gamma(\tilde{\beta}; \mathbf{X})\right]\right)\right| \\
&= \frac{1}{n\alpha}\left|\log\left(\sum_{\beta \in \tilde{\mathcal{S}}_\delta}\exp\left[-n\alpha\left(\hat{R}_n^\gamma(\beta; \mathbf{X}) - \hat{R}_n^\gamma(\tilde{\beta}; \mathbf{X})\right)\right]\right)\right| \\
&< \frac{1}{n\alpha}\log\left|\tilde{\mathcal{S}}_\delta\right|, \qquad (44)
\end{aligned}
$$

where the last line follows from the fact that $\hat{R}_n^\gamma(\beta; \mathbf{X}) - \hat{R}_n^\gamma(\tilde{\beta}; \mathbf{X})$ is always non-negative for $\beta \in \tilde{\mathcal{S}}_\delta$ by definition, and is zero at least once, meaning when we multiply by $-n\alpha$ they will always be either zero or strictly negative. This means the sum inside of the logarithm lies in $\left(1, \left|\tilde{\mathcal{S}}_\delta\right|\right)$. By Proposition 4.2.12 of Vershynin (2018), we can say that since $\mathcal{S}_p \subseteq B_{\mathbb{R}^p}(\mathbf{0}, \mathsf{L}\sqrt{p})$, then

$$
\left|\tilde{\mathcal{S}}_\delta\right| \leq \left(\frac{3\mathsf{L}\sqrt{p}}{\delta\sqrt{p}}\right)^p = \left(\frac{3\mathsf{L}}{\delta}\right)^p,
$$

and so combining this with (44) we have

$$
\mathbb{D}_2 \leq \frac{p}{n\alpha}\log\left(\frac{3\mathsf{L}}{\delta}\right) \leq \mathsf{C}_3\frac{1}{\alpha}\log\left(\frac{1}{\delta}\right)
$$

for n sufficiently large. We finish by noting that $\mathbb{D}_4$ and $\mathbb{D}_5$ have the exact same bounds as $\mathbb{D}_2$ and $\mathbb{D}_1$, respectively. ■

H.3. Main lemmas under independence under partial block independence.

Let $i \leq n$ be an index. Throughout this subsection we will assume that the block $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j \in \mathcal{B}_i}$ is independent from $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j \notin \mathcal{B}_i}$ where $\mathcal{B}_i$ is the block of size k that contains i.

H.3.1. MAIN LEMMA

In the next result, we recall the notation that $\tilde{\mathbf{U}}_i \equiv \tilde{\mathbf{U}}_i(t) = \cos(t)\mathbf{X}_i - \sin(t)\mathbf{G}_i$ for $t \in [0, 1]$, where we abbreviate the dependence on t for the simplicity of presentation in the proof.

Lemma 24 *Let $i \leq n$. Let $(\mathbf{X}_j, y_j(\mathbf{X}_j))$ and $(\mathbf{G}_j, y_j(\mathbf{G}_j))$ be generated under Assumptions 2-5 and that* $\mathrm{Var}(\mathbf{G}) = \mathrm{Var}(\mathbf{X})$. *In addition, suppose that the block $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j \in \mathcal{B}_i}$ is independent from $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j \notin \mathcal{B}_i}$. For every $t \in (0, \frac{\pi}{2})$, we have*

$$\limsup_{n \to \infty} \left| \mathbb{E}\left[-h'\left(f_\delta(\mathbf{U})\right) \langle \tilde{U}_i^\top \mathcal{D}_i(\beta) \rangle \right] \right| \leq \tau \tag{45}$$

and

$$\limsup_{n \to \infty} \left| \mathbb{E}\left[-h'\left(f_\delta(\mathbf{U})\right) \langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle \right] \right| \leq \tau \tag{46}$$

where we defined $\mathcal{D}_{i,L}(\beta) := \mathcal{D}_i(\beta) \mathbb{I}(|\tilde{U}_i^T \beta|, |\tilde{U}_i^T \beta| \leq L)$

Proof We note that it is enough to show the desired result for $\mathcal{D}_{i,L}$ as Eq. (45) directly follows by taking $L = \infty$.

Firstly by adding and subtracting the quantity $h'(f_\delta(\mathbf{U}^{ik}))$ we obtain

$$\begin{aligned} &\left| \mathbb{E}\left[-h'\left(f_\delta(\mathbf{U})\right) \langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle \right] \right| \\ &\leq \underbrace{\mathbb{E}\left| \left(h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik})) \right) \langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle \right|}_{(a)} + \underbrace{\left| \mathbb{E}\left[h'(f_\delta(\mathbf{U}^{ik})) \langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle \right] \right|}_{(b)}. \end{aligned}$$

For the term (a), we use Cauchy-Schwarz to say that

$$\mathbb{E}\left| \left(h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik})) \right) \langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle \right| \leq \underbrace{\sqrt{\mathbb{E}\left[h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik})) \right]^2}}_{(a_1)} \cdot \underbrace{\sqrt{\mathbb{E}\langle \tilde{U}_i^\top \mathcal{D}_{i,L}(\beta) \rangle^2}}_{(a_2)}.$$

To control these two terms, we first apply Lemma 25 to (a_1) to obtain

$$(a_1) \leq \sqrt{\frac{\mathsf{C}_1 k^2}{n}} = \frac{\mathsf{C}_2 k}{\sqrt{n}}, \tag{47}$$

and then apply Lemma 26 to (a_2) to say that

$$
\begin{aligned}
\mathbb{E}\langle \tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i,L}(\beta)\rangle^2 &\overset{(i)}{=} \mathbb{E}\left\langle \frac{\tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i,L}(\beta) e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}}{\langle e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}\rangle_{i,k}} \right\rangle_{i,k}^2 \\
&\overset{(ii)}{\leq} \mathbb{E}\left\langle \left(\frac{\tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i}(\beta) e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}}{\langle e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}\rangle_{i,k}} \right)^2 \right\rangle_{i,k} \\
&= \mathbb{E}\left\langle \mathbb{E}_{(i,k)} \left[\left(\frac{\tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i}(\beta) e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}}{\langle e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}\rangle_{i,k}} \right)^2 \right] \right\rangle_{i,k} \\
&\overset{(iii)}{\leq} \mathbb{E}\left[\sup_{\beta\in\tilde{\mathcal{S}}_\delta} \mathbb{E}_{(i,k)} \left[\left(\frac{\tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i}(\beta) e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}}{\langle e^{-\alpha \sum_{j\in\mathcal{B}_i} \ell_j(\beta)}\rangle_{i,k}} \right)^2 \right] \right] \\
&\overset{(iv)}{\leq} \mathsf{C}_1(k,\alpha,\gamma),
\end{aligned} \tag{48}
$$

where (i) comes from noting that, for a general function g,

$$
\begin{aligned}
\langle \mathsf{g}(\beta)\rangle = \sum_{\beta\in\tilde{\mathcal{S}}_\delta} w_\gamma(\beta)\mathsf{g}(\beta) = \frac{\sum_{\beta\in\tilde{\mathcal{S}}_\delta} e^{-n\alpha \hat{R}_n^\gamma(\beta)}\mathsf{g}(\beta)}{\sum_{\beta'\in\tilde{\mathcal{S}}_\delta} e^{-n\alpha\hat{R}_n^\gamma(\beta')}} &= \frac{\sum_{\beta\in\mathcal{S}_\delta} e^{-n\alpha\hat{R}_n^{\gamma,i,k}(\beta)}\mathsf{g}(\beta)e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}}{\sum_{\beta'\in\tilde{\mathcal{S}}_\delta} e^{-n\alpha\hat{R}_n^\gamma(\beta')}e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta')}} \\
&= \left\langle \frac{\mathsf{g}(\beta)e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}} \right\rangle_{i,k},
\end{aligned} \tag{49}
$$

(ii) follows from Jensen's Inequality combined with the fact that $|\tilde{U}_i^T\mathcal{D}_{i,L}(\beta)| \overset{a.s}{\leq} \tilde{U}_i^T\mathcal{D}_i(\beta)$, (iii) from the fact that

$$
\langle \mathsf{g}(\beta)\rangle \leq \sup_{\beta\in\tilde{\mathcal{S}}_\delta} g(\beta), \tag{50}
$$

and (iv) is exactly the statement of Lemma 26 for some $\mathsf{C}_1(k,\alpha,\gamma)$. Combining (47) and (48), we conclude that

$$
(a) \leq \frac{\mathsf{C}_2 k \mathsf{C}_1(k,\alpha,\gamma)}{\sqrt{n}} =: \frac{\mathsf{C}_2(k,\alpha,\gamma)}{\sqrt{n}}. \tag{51}
$$

For the term (b), we first apply Lemma 27, which says that there exist $\mathsf{D} = \mathsf{D}(k,\alpha,\gamma,\tau)$ and real coefficients $b_0,\ldots,b_{\mathsf{D}}$ such that

$$
(b) \leq \tau + \sum_{\ell=0}^{\mathsf{D}} |b_\ell| \sup_{\substack{\beta_0,\ldots,\beta_\ell \\ \in\tilde{\mathcal{S}}}} \left| \mathbb{E}\left[\tilde{U}_i^{\mathsf{T}} \mathcal{D}_{i,L}(\beta_0) \exp\Big(-\alpha \sum_{r=0}^{\ell} \sum_{j\in\mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}}\beta_r) \Big) \right] \right|, \tag{52}
$$

noting by definition that $|b_\ell| = \binom{\mathsf{D}}{\ell} \leq \mathsf{D}!$ is bounded. Now, we define the Gaussian interpolator

$$
\mathbf{V} := \sin(t)\tilde{\mathbf{G}} + \cos(t)\mathbf{G}
$$

with $\tilde{\mathbf{G}}$ as defined in Lemma 28 being an identical copy of $\mathbf{G}$. Moreover note that according to Theorem 40 for all $\epsilon > 0$ there exists $g_\epsilon(\cdot)$ such that $\|g_\epsilon\|_\infty \leq 1$ and $\|g_\epsilon\|_{\text{Lipchitz}} \leq 2\epsilon^{-1}$ and such that

$$\Big|g_\epsilon(|\tilde{U}_i^T\beta|, |U_i^T\beta|) - \mathbb{I}(|\tilde{U}_i^T\beta|, |U_i^T\beta| \leq L)\Big| \leq \mathbb{I}\Big[|\tilde{U}_i^T\beta| \in [L-\epsilon, L]\Big] + \mathbb{I}\Big[|U_i^T\beta| \in [L-\epsilon, L]\Big].$$

We note that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\substack{\beta_0,\dots,\beta_\ell \\ \in\tilde{S}}} \Bigg|\mathbb{E}\Bigg[\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0) g_\epsilon(|\tilde{U}_i^T\beta_0|, |U_i^T\beta_0|) \exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j, U_j^\top\beta_r)\Big)\Bigg] \\
&\qquad - \mathbb{E}\Bigg[\tilde{U}_i^\top \mathcal{D}_{i,L}(U_{\mathcal{B}_i},\beta_0) \exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j, U_j^\top\beta_r)\Big)\Bigg]\Bigg| \\
&\leq \limsup_{n\to\infty} \sup_{\beta_0\in\tilde{S}} \mathbb{E}\Big[\Big|\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)\Big| \mathbb{I}(|U_i^T\beta_0| \in [L-\epsilon, L])\Big] \\
&\qquad + \mathbb{E}\Big[\Big|\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)\Big| \mathbb{I}(|\tilde{U}_i^T\beta_0| \in [L-\epsilon, L])\Big] \\
&\leq \limsup_{n\to\infty} \sup_{\beta_0\in\tilde{S}} \mathbb{E}\Big[\Big|\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)\Big|^2\Big]^{1/2} \sqrt{P(|U_i^T\beta_0| \in [L-\epsilon, L]) + P(|\tilde{U}_i^T\beta_0| \in [L-\epsilon, L])}
\end{aligned}$$

Using assumption 3 we know that

$$\limsup_{n\to\infty} \sup_{\beta_0\in\tilde{S}} \mathbb{E}\Big[\Big|\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)\Big|\Big] < \infty.$$

Moreover we note that

$$\begin{aligned}
P(|U_i^T\beta_0| \in [L-\epsilon, L]) &= \mathbb{E}\Big(\Phi^c\Big(\frac{L-\epsilon-\sin(\mathrm{t})\mathrm{X}_\mathrm{i}^\mathrm{T}\beta_0}{\cos(\mathrm{t})}\Big)\Big) - \mathbb{E}\Big(\Phi^c\Big(\frac{L-\sin(\mathrm{t})\mathrm{X}_\mathrm{i}^\mathrm{T}\beta_0}{\cos(\mathrm{t})}\Big)\Big) \\
&\leq \frac{\epsilon}{\cos(\mathrm{t})}\mathbb{E}\Big(\varphi\Big(\frac{L-\epsilon-\sin(\mathrm{t})\mathrm{X}_\mathrm{i}^\mathrm{T}\beta_0}{\cos(\mathrm{t})}\Big)\Big) \\
&\leq \frac{\epsilon}{\cos(\mathrm{t})}
\end{aligned}$$

Similarly we can obtain that

$$P(|\tilde{U}_i^T\beta_0| \in [L-\epsilon, L]) \leq \frac{\epsilon}{\sin(\mathrm{t})}.$$

Hence we obtain that there is a constant $\mathsf{C}_3 > 0$ such that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\substack{\beta_0,\dots,\beta_\ell \\ \in\tilde{S}}} \Bigg|\mathbb{E}\Bigg[\tilde{U}_i^\top \mathcal{D}_i(U_{\mathcal{B}_i},\beta_0) g_\epsilon(|\tilde{U}_i^T\beta_0|, |U_i^T\beta_0|) \exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j, U_j^\top\beta_r)\Big)\Bigg] && (53) \\
&\qquad - \mathbb{E}\Bigg[\tilde{U}_i^\top \mathcal{D}_{i,L}(U_{\mathcal{B}_i},\beta_0) \exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j, U_j^\top\beta_r)\Big)\Bigg]\Bigg| && (54) \\
&\leq \mathsf{C}_3\sqrt{\epsilon}. && (55)
\end{aligned}$$

Moreover we note that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\substack{\beta_0,\dots,\beta_\ell \\ \in \tilde{\mathcal{S}}}} \Bigg|\mathbb{E}\Bigg[\tilde{U}_i^{\mathsf{T}}\mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)g_\epsilon(|\tilde{U}_i^T\beta_0|,|U_i^T\beta_0|)\exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^{\mathsf{T}}\beta_r)\Big)\Bigg]\Bigg]\Bigg| \\
&\overset{(i)}{\le} \limsup_{n\to\infty} \sup_{\substack{\beta_0,\dots,\beta_\ell \\ \in \tilde{\mathcal{S}}}} \Bigg|\mathbb{E}\Bigg[\tilde{V}_i^{\mathsf{T}}\mathcal{D}_i(V_{\mathcal{B}_i},\beta_0)g_\epsilon(|\tilde{V}_i^T\beta_0|,|V_i^T\beta_0|)\exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,V_j^{\mathsf{T}}\beta_r)\Big)\Bigg]\Bigg]\Bigg| \\
&= \limsup_{n\to\infty} \sup_{\substack{\beta_0,\dots,\beta_\ell \\ \in \tilde{\mathcal{S}}}} \Bigg|\mathbb{E}\Bigg[\mathbb{E}\Big(g_\epsilon(|\tilde{V}_i^T\beta_0|,|V_i^T\beta_0|)\tilde{V}_i\Big|(V_j)\Big)^{\mathsf{T}}\mathcal{D}_i(V_{\mathcal{B}_i},\beta_0)\exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,V_j^{\mathsf{T}}\beta_r)\Big)\Bigg]\Bigg]\Bigg| \\
&\overset{(ii)}{=} 0. \qquad (56)
\end{aligned}$$

Above, (i) follows from the second statement of Lemma 28 coupled with Assumption 5, since we know that the function

$$g(X_{\mathcal{B}_i}B, G_{\mathcal{B}_i}B) := \tilde{U}_i^{\mathsf{T}}\mathcal{D}_i(U_{\mathcal{B}_i},\beta_0)g_\epsilon(|\tilde{V}_i^T\beta_0|,|V_i^T\beta_0|)\exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^{\mathsf{T}}\beta_r)\Big),$$

where $B := (\beta_0,\dots,\beta_\ell) \in \mathbb{R}^{p\times(\ell+1)}$ is locally Lipschitz as all of its components are, and is also square integrable as

$$\sup_B \mathbb{E}\Big[g(X_{\mathcal{B}_i}B, G_{\mathcal{B}_i}B)^2\Big] \le \sup_B \mathbb{E}\Big[\Big(\tilde{U}_i^{\mathsf{T}}\mathcal{D}_i(\beta_0)\Big)^2\Big] \le \mathsf{C}(k,\gamma) \tag{57}$$

by the argument in (59) and (61) of Lemma 26. Further, (ii) follows from independence of $\tilde{V}_i$ from (V_j), as

$$E[\tilde{V}_iV_j^{\mathsf{T}}] = \sin(t)\cos(t)\mathbb{E}[X_iX_j^{\mathsf{T}}] - \sin(t)\cos(t)\mathbb{E}[G_iG_j^{\mathsf{T}}] = 0,$$

and we know that two zero-covariance Gaussians are necessarily independent of each other, combined with the fact that $x \to g_\epsilon(x^T\beta_0, V_i^T\beta_0)x$ is a symmetric function. We may thus combine (51), (52), (53) and (56) to conclude

$$\limsup_{n\to\infty}\Big|\mathbb{E}\Big[-h'\left(f_\delta(\mathbf{U})\right)\langle\tilde{U}_i^{\mathsf{T}}\mathcal{D}_{i,L}(\beta)\rangle\Big]\Big| \le \limsup_{n\to\infty}\frac{\mathsf{C}_2(k,\alpha,\gamma)}{\sqrt{n}} + \tau + \mathsf{C}_3\sqrt{\epsilon} = \tau + \mathsf{C}_3\sqrt{\epsilon}, \tag{58}$$

As the dependence on ϵ is arbitrary we get the desired result. ■

H.3.2. UPPER BOUNDS ON EXPECTATIONS & APPROXIMATIONS

The next set of lemmas are used to bound various expectations that appear in the proofs of our main results. Throughout this subsection $i \le n$ designates an index. We assume that $(\mathbf{X}_j, y_j(\mathbf{X}_j))$ and $(\mathbf{G}_j, y_j(\mathbf{G}_j))$ are generated under Assumptions 2-5 and that $\mathrm{Var}(\mathbf{G}) = \mathrm{Var}(\mathbf{X})$. In addition, suppose that the block $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j\in\mathcal{B}_i}$ is independent from $(\mathbf{X}_j, y_j(\mathbf{X}_j))_{j\notin\mathcal{B}_i}$. We begin with a Lindeberg bound.

Lemma 25 (Lindeberg Difference Bound) *Let $\alpha, \delta > 0$. Then there exists $\mathsf{C} > 0$ such that for each $t \in [0, \frac{\pi}{2}]$ and n sufficiently large,*

$$\mathbb{E}\left(\left[h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik}))\right]^2\right) \leq \frac{\mathsf{C}k^2}{n}.$$

Proof By the Lipschitzness of h', we first have

$$\begin{aligned}
\left|h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik}))\right| &\leq \|h'\|_{\mathrm{Lip}}\left|f_\delta(\mathbf{U}) - f_\delta(\mathbf{U}^{ik})\right| \\
&\overset{(i)}{\leq} \frac{\mathsf{C}_1}{n\alpha}\left|\log\left(\frac{\sum_\beta \exp\left[-n\alpha\hat{R}_n^\gamma(\beta; \boldsymbol{\eta}, \mathbf{U})\right]}{\sum_\beta \exp\left[-n\alpha\hat{R}_n^\gamma(\beta; \boldsymbol{\eta}, \mathbf{U}^{ik})\right]}\right)\right| \\
&\overset{(ii)}{\leq} \frac{\mathsf{C}_1}{n\alpha}\left(k\alpha\log(2) + \left\langle \alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\right\rangle_{i,k}\right) \\
&\leq \frac{\mathsf{C}_1}{n}\left(k + \left\langle \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\right\rangle_{i,k}\right),
\end{aligned}$$

where (i) is because $h \in \mathcal{H}$, and (ii) is via Jensen's inequality on the natural logarithm and the fact that $\ell(a, 0) = \log(2)$ for any $a \in \mathbb{R}$. Thus, we have

$$\mathbb{E}\left[h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik}))\right]^2 \leq \frac{\mathsf{C}_1}{n^2}\left(\mathbb{E}\left\langle \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\right\rangle_{i,k}^2 + k\mathbb{E}\left\langle \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\right\rangle_{i,k} + k^2\right).$$

From here, we can first notice that

$$\log(1 + e^{-x}) \leq |x| + 1 \qquad \text{for all } x \in \mathbb{R},$$

and thus, using $\left|\eta_j\right| = 1, \omega_j \leq 1$, and Cauchy-Schwarz, we can say that

$$\begin{aligned}
\sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta) &\leq \sum_{j \in \mathcal{B}_i}\left(\left|\eta_j U_j^\intercal \beta\right| + 1\right) \\
&\leq k + \|\beta\| \sum_{j \in \mathcal{B}_i} \|U_j\|,
\end{aligned}$$

and so

$$
\begin{aligned}
\mathbb{E}\Big\langle \sum_{j\in\mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\Big\rangle_{i,k} &\le k + \sum_{j\in\mathcal{B}_i} \mathbb{E}\Big\langle \|\beta\|\,\|U_j\|\Big\rangle_{i,k} \\
&\le k + \sum_{j\in\mathcal{B}_i} \mathbb{E}\left[\|U_j\| \sup_{\beta\in\tilde{\mathcal{S}}_\delta} \|\beta\|\right] \\
&\le k + \mathsf{L}\sqrt{p} \sum_{j\in\mathcal{B}_i} \mathbb{E}\left[\|U_j\|\right] \\
&\overset{(i)}{\le} k + \mathsf{L}\sqrt{p} \sum_{j\in\mathcal{B}_i} \sqrt{\sum_{k=1}^{p} \mathbb{E}[U_{jk}^2]} \\
&\overset{(ii)}{\le} k + \mathsf{L}k\sqrt{p} \sup_j \sqrt{\mathrm{Tr}(\Sigma_j)} \\
&\overset{(iii)}{\le} \mathsf{C}_2 k\sqrt{p},
\end{aligned}
$$

where (i) is via Jensen's Inequality, (ii) is because

$$
\mathbb{E}\left[U_{jk}^2\right] = \mathbb{E}\left[\sin^2(t)X_{jk}^2 + \cos^2(t)G_{jk}^2 + \sin(t)\cos(t)X_{jk}G_{jk}\right] = (\Sigma_j)_{k,k}\Big(\sin^2(t) + \cos^2(t)\Big) = (\Sigma_j)_{k,k},
$$

and (iii) is via Lemma 44 and the fact that

$$
\mathrm{Tr}(\Sigma_j) = \sum_{i=1}^{p} \langle \Sigma_j e_i, e_i\rangle \le \sum_{i=1}^{p} \|\Sigma_j\|_{\mathrm{op}} \le p\|\Sigma_j\|_{\mathrm{op}} = O(1),
$$

where the last inequality comes from the scaling of $\mathbf{U}$. Using an identical argument and the fact that

$$
\mathbb{E}\Big\langle \sum_{j\in\mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\Big\rangle_{i,k}^2 \le \mathbb{E}\Big\langle\Big(\sum_{j\in\mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\Big)^2\Big\rangle_{i,k},
$$

we similarly obtain

$$
\mathbb{E}\Big\langle \sum_{j\in\mathcal{B}_i} \omega_j \ell(\eta_j, U_j^\intercal \beta)\Big\rangle_{i,k}^2 \le \mathsf{C}_3 k^2 p,
$$

and thus

$$
\mathbb{E}\left(\left[h'(f_\delta(\mathbf{U})) - h'(f_\delta(\mathbf{U}^{ik}))\right]^2\right) \le \frac{\mathsf{C}_1}{n^2}\left(\mathsf{C}_3 k^2 p + \mathsf{C}_2 k^2\sqrt{p} + k^2\right) \le \frac{\mathsf{C}k^2}{n}
$$

for n sufficiently large. ■

Lemma 26 (Second Moment Bound) *There exists $\mathsf{C}(k,\alpha,\gamma) > 0$ such that for every $t \in \left[0, \frac{\pi}{2}\right]$ and $i = 1, \ldots, n$, we have*

$$
\sup_{\beta\in\tilde{\mathcal{S}}} \mathbb{E}_{(i,k)}\left[\left(\frac{\tilde{U}_i^\intercal \mathcal{D}_i(\beta) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}}\right)^2\right] \le \mathsf{C}(k,\alpha,\gamma).
$$

Proof Fix some $\beta \in \tilde{\mathcal{S}}$. We may bound

$$\begin{aligned}
\mathbb{E}_{(i,k)}\left[\left(\frac{\tilde{U}_i^\top \mathcal{D}_i(\beta) e^{-\alpha \sum_{j\in\mathcal{B}_i} \omega_j \ell_j(\beta)}}{\langle e^{-\alpha \sum_{j\in\mathcal{B}_i} \omega_j \ell_j(\beta)} \rangle_{i,k}}\right)^2\right] &\overset{(i)}{\leq} \mathbb{E}_{(i,k)}\left[\left(\tilde{U}_i^\top \mathcal{D}_i(\beta)\right)^4\right]^{1/2} \mathbb{E}_{(i,k)}\left[\left(\frac{e^{-\alpha \sum_j \omega_j \ell_j(\beta)}}{\langle e^{-\alpha \sum_j \omega_j \ell_j(\beta)} \rangle_{i,k}}\right)^4\right]^{1/2} \\
&\overset{(ii)}{\leq} \mathbb{E}\left[\left(\tilde{U}_i^\top \mathcal{D}_i(\beta)\right)^4\right]^{1/2} \left\langle \mathbb{E}\left[e^{4\alpha \sum_j \omega_j \ell_j(\beta)}\right] \right\rangle_{i,k}^{1/2} \\
&\leq \underbrace{\mathbb{E}\left[\left(\tilde{U}_i^\top \mathcal{D}_i(\beta)\right)^4\right]^{1/2}}_{(a)} \underbrace{\left(\sup_\beta \mathbb{E}\left[e^{4\alpha \sum_j \omega_j \ell_j(\beta)}\right]\right)^{1/2}}_{(b)},
\end{aligned}$$

where (i) is via Cauchy-Schwarz, and (ii) is from the fact that $\mathbb{E}_{(i,k)}[\ \cdot\ ]$ and $\langle\ \cdot\ \rangle_{i,k}$ commute due to $w_\gamma^{i,k}(\beta)$ only being a function of $(U_j)_{j\notin\mathcal{B}_i}$, and from applying Jensen's Inequality to the convex function x^{-4}, using the fact that

$$\alpha \sum_{j\in\mathcal{N}_i} \omega_j \ell_j(\beta) \geq 0 \implies \exp\left(-\alpha \sum_{j\in\mathcal{N}_i} \omega_j \ell_j(\beta)\right) \leq 1.$$

Also in (ii), note that the conditional expectation has disappeared due to block dependence. For the term (a), we can notice that

$$\begin{aligned}
\left|\tilde{U}_i^\top \mathcal{D}_i(\beta)\right| &= \left|\eta_i \omega_i \sigma_{i\beta} \tilde{U}_i^\top \beta + \sum_{j\in\mathcal{B}_i} \omega_j \sigma_{j\beta} \eta_j' a_{ji} U_j^\top \beta \tilde{U}_i^\top \beta^*\right| \\
&\overset{(i)}{\leq} \left|\tilde{U}_i^\top \beta\right| + \frac{3}{\gamma}\left|\tilde{U}_i^\top \beta^*\right| \sum_{j\in\mathcal{B}_i} \left|U_j^\top \beta\right|, \qquad (59)
\end{aligned}$$

where (i) is from $\eta_i, \omega_i, \sigma_{i\beta}, a_{ji} \leq 1$ and Lemma 21 bounding $\left|\eta_j'\right|$. From here, we know that since the rows of $\mathbf{X}$ and $\mathbf{G}$ are sub-Gaussian by Assumption 3, so must those of $\mathbf{U}$ and $\tilde{\mathbf{U}}$ be as well, since

$$\|U_i\|_{\psi_2} = \|\sin(t)X_i + \cos(t)G_i\|_{\psi_2} \leq |\sin(t)|\|X_i\|_{\psi_2} + |\cos(t)|\|G_i\|_{\psi_2} \leq \frac{\sqrt{2}\mathsf{K}_X}{\sqrt{n}},$$

and similarly for $\tilde{\mathbf{U}}$. This further implies that for any $\beta \in \mathcal{S}_p$ we have

$$\begin{aligned}
\|U_i^\top \beta\|_{\psi_2} &\leq \frac{\sqrt{2}\mathsf{K}_X \|\beta\|}{\sqrt{n}} \\
&\leq \frac{\sqrt{2}\mathsf{K}_X \mathsf{L}\sqrt{p}}{\sqrt{n}} \\
&\leq \mathsf{C}_1 \qquad (60)
\end{aligned}$$

for n sufficiently large, and again this holds for $\tilde{U}_i^\intercal \beta$ as well. We conclude by a multinomial expansion that

$$
\begin{aligned}
\mathbb{E}\left[\left(\tilde{U}_i^\intercal \mathcal{D}_i(\beta)\right)^4\right] &\overset{(i)}{\leq} \sum_{\ell=0}^{4}\binom{4}{\ell}\mathbb{E}\left[\left|\tilde{U}_i^\intercal \beta\right|^{4-\ell}\frac{3^\ell}{\gamma^\ell}\left|\tilde{U}_i^\intercal \beta^*\right|^{\ell}\left(\sum_{j\in\mathcal{B}_i}\left|U_j^\intercal \beta\right|\right)^{\ell}\right] \\
&\overset{(ii)}{\leq} \frac{486}{\gamma^4}\sum_{\ell=0}^{4}\mathbb{E}\left[\left|\tilde{U}_i^\intercal \beta\right|^{4-\ell}\left|\tilde{U}_i^\intercal \beta^*\right|^{\ell}\left(\sum_{j\in\mathcal{B}_i}\left|U_j^\intercal \beta\right|\right)^{\ell}\right] \\
&\overset{(iii)}{\leq} \frac{162}{\gamma^4}\sum_{\ell=0}^{4}\mathbb{E}\left[\left|\tilde{U}_i^\intercal \beta\right|^{12-3\ell} + \left|\tilde{U}_i^\intercal \beta^*\right|^{3\ell} + \left(\sum_{j\in\mathcal{B}_i}\left|U_j^\intercal \beta\right|\right)^{3\ell}\right] \\
&\overset{(iv)}{\leq} \frac{162}{\gamma^4}\sum_{\ell=0}^{4} l\left(\mathsf{C}_2\sqrt{12-3\ell}\right)^{12-3\ell} + \left(\mathsf{C}_3\sqrt{3\ell}\right)^{3\ell} + \left(\mathsf{C}_4 k\sqrt{3\ell}\right)^{3\ell} \\
&\overset{(v)}{\leq} \mathsf{C}_5\gamma^{-4}k^{12},
\end{aligned}
\tag{61}
$$

where (i) is from (59) and binomial expansion, (ii) is because $\ell \leq 4$ and $\max_\ell \binom{4}{\ell} = 6$, (iii) is the AM-GM Inequality for $n = 3$, (iv) is from Proposition 2.5.2 in Vershynin (2018) on equivalent properties of sub-Gaussian random variables, and (v) is from the fact that

$$3\ell \vee (12-3\ell) \leq 12 \quad \text{for all} \quad 0 \leq \ell \leq 4.$$

For the term (b), we once again use that $|\omega_i| \leq 1$ and $\log(1+e^{-x}) \leq |x|+1$ for $x \in \mathbb{R}$ to say that

$$
\begin{aligned}
\mathbb{E}\left[e^{4\alpha\sum_j \omega_j \ell_j(\beta)}\right] &\leq \mathbb{E}\left[e^{4\alpha\sum_j |U_j^\intercal \beta|+1}\right] \\
&\overset{(i)}{\leq} e^{4k\alpha}\mathbb{E}\left[e^{4\alpha\sum_j |U_j^\intercal \beta|}\right] \\
&\leq e^{4k\alpha}\mathbb{E}\left[\prod_{j\in\mathcal{B}_i} e^{4\alpha|U_j^\intercal \beta|}\right] \\
&\overset{(ii)}{\leq} e^{4k\alpha}\left(\prod_{j\in\mathcal{B}_i}\mathbb{E}\left[e^{4k\alpha|U_j^\intercal \beta|}\right]\right)^{1/k} \\
&\overset{(iii)}{\leq} e^{4k\alpha(\mu+1)}\left(\prod_{j\in\mathcal{B}_i}\mathbb{E}\left[e^{4k\alpha(|U_j^\intercal \beta|-\mu)}\right]\right)^{1/k} \\
&\overset{(iv)}{\leq} e^{4k\alpha(\mu+1)}e^{\mathsf{C}_1 k^2 \mathsf{K}_X^2\alpha^2} \\
&\leq e^{\mathsf{C}_2 k^2\alpha^2},
\end{aligned}
\tag{62}
$$

where (i) is via $|\mathcal{B}_i| = k$, (ii) is via Hölder's Inequality, (iii) is from adding and subtracting $\mu := \mathbb{E}[|U_j^\intercal \beta|]$ in the exponent, which satisfies

$$\mu \leq \sqrt{\mathbb{E}\left[(U_j^\intercal \beta)^2\right]} = \sqrt{\beta^\intercal \Sigma_j \beta} \leq \|\beta\|\,\|\Sigma_j\|_{\mathrm{op}}^{1/2} \leq \frac{\mathsf{C}\mathsf{K}_X\mathsf{L}\sqrt{p}}{\sqrt{n}} \leq \mathsf{C}_3$$

for n sufficiently large, by Jensen's Inequality and Lemma 44, and (iv) is via sub-Gaussianity of the centered version of $U_j^\top\beta$, which is sub-Gaussian by Lemma 2.6.8 of Vershynin (2018), and thus satisfies Condition (v) of Proposition 2.5.2 of the same text. Since this holds for all $\beta \in \tilde{\mathcal{S}}$, we conclude by combining (61) and (62) that

$$\sup_{\beta\in\tilde{\mathcal{S}}} \mathbb{E}_{(i,k)}\left[\left(\frac{\tilde{U}_i^\top \mathcal{D}_i(\beta) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}}\right)^2\right] \le \mathsf{C}_5\gamma^{-2}k^6\exp\left(\mathsf{C}_2k^2\alpha^2\right) =: \mathsf{C}(k,\alpha,\gamma).$$

■

The following lemma employs a technique developed in Montanari and Saeed (2022), which will allow us to convert a complicated term involving the inverse function $1/x$ into one involving a polynomial that is much more straightforward to control.

Lemma 27 (Polynomial Approximation) *Let $\alpha, \delta, \gamma, \tau > 0$. Then there exists $\mathsf{D} = \mathsf{D}(k,\alpha,\gamma,\tau)$ and coefficients $b_0,\dots,b_{\mathsf{D}} \in \mathbb{R}$ such that*

$$\begin{aligned}&\left|\mathbb{E}\left[h'(f_\delta(\mathbf{U}^{ik}))\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle\right]\right| \\ &\qquad\le \tau + \sum_{\ell=0}^{\mathsf{D}}|b_\ell| \sup_{\substack{\beta_0,\dots,\beta_\ell\\ \in\tilde{\mathcal{S}}}}\left|\mathbb{E}\left[\tilde{U}_i^\top\mathcal{D}_i(\beta_0)\exp\Big(-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta_r)\Big)\right]\right|.\end{aligned}$$

Proof Since $\mathbf{U}^{ik}$ was constructed to be independent of $U_{\mathcal{B}_i}$ and $\tilde{U}_{\mathcal{B}_i}$, we first expand this quantity as

$$\begin{aligned}&\left|\mathbb{E}\left[h'(f_\delta(\mathbf{U}^{ik})\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle\right]\right| \\ &\overset{(i)}{=}\left|\mathbb{E}\left[\mathbb{E}_{(i,k)}\left[h'(f_\delta(\mathbf{U}^{ik})\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle\right]\right]\right| \\ &\overset{(ii)}{=}\left|\mathbb{E}\left[h'(f_\delta(\mathbf{U}^{ik})\mathbb{E}_{(i,k)}\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle\right]\right| \\ &\le \|h'\|_\infty\cdot\mathbb{E}\left|\mathbb{E}_{(i,k)}\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle\right| \\ &\overset{(iii)}{\le}\mathbb{E}\left|\mathbb{E}_{(i,k)}\left\langle\frac{\tilde{U}_i^\top\mathcal{D}_i(\beta)e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta)}\rangle_{i,k}}\right\rangle_{i,k}\right| \\ &\overset{(iv)}{=}\mathbb{E}\left|\left\langle\mathbb{E}_{(i,k)}\left(\frac{\tilde{U}_i^\top\mathcal{D}_i(\beta)e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta)}\rangle_{i,k}}\right)\right\rangle_{i,k}\right| \\ &\overset{(v)}{\le}\mathbb{E}\left[\sup_{\beta_0\in\tilde{\mathcal{S}}}\underbrace{\left|\mathbb{E}_{(i,k)}\left(\frac{\tilde{U}_i^\top\mathcal{D}_i(\beta_0)e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta_0)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\top\beta)}\rangle_{i,k}}\right)\right|}_{(a)}\right],\end{aligned} \tag{63}$$

where (i) is via the law of total expectation, (ii) is by the independence mentioned above, (iii) is by definition of $h \in \mathcal{H}$ and the ability to rewrite $\langle\tilde{U}_i^\top\mathcal{D}_i(\beta)\rangle$ as done previously in (49), (iv) is because $\mathbb{E}_{(i,k)}[\,\cdot\,]$ and $\langle\,\cdot\,\rangle_{i,k}$ commute with each other, and (v) is the same as in (50).

Now, we will approximate the inverse function x^{-1} by a polynomial by defining the functions

$$Q_{\mathsf{D}}(x) := \sum_{\ell=0}^{\mathsf{D}} (1-x)^{\ell} = \sum_{\ell=0}^{\mathsf{D}} b_\ell x^\ell, \qquad R_{\mathsf{D}}(x) := \frac{1}{x} - Q_{\mathsf{D}}(x)$$

for some degree D and $x \in (0,1]$. Recall from Lemma 26 that there exists $\mathsf{C}(k,\alpha,\gamma)$ such that

$$\sup_{\beta\in\tilde{\mathcal{S}}} \mathbb{E}_{(i,k)}\left[\left(\frac{\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)}}{\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}}\right)^2\right] \leq \mathsf{C}(k,\alpha,\gamma). \tag{64}$$

Thus we may choose the degree of our polynomial to be the exact $\mathsf{D} = \mathsf{D}\left(k,\alpha,\tau^2/\mathsf{C}(k,\alpha,\gamma)\right)$ such that, by Lemma 43, we have

$$\mathbb{E}_{(i,k)}\left[R_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\rangle_{i,k}\right)^2\right] < \frac{\tau^2}{\mathsf{C}(k,\alpha,\gamma)}. \tag{65}$$

We conclude that

$$\begin{aligned}
(a) &\overset{(i)}{\leq} \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} Q_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)\right]\right| \\
&\qquad + \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} R_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)\right]\right| \\
&\overset{(ii)}{\leq} \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} Q_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)\right]\right| \\
&\qquad + \mathbb{E}_{(i,k)}\left[\left(\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)}\right)^2\right]^{1/2} \mathbb{E}_{(i,k)}\left[R_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)^2\right]^{1/2} \\
&\overset{(iii)}{\leq} \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} Q_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)\right]\right| + \tau,
\end{aligned} \tag{66}$$

where (i) is via the Triangle Inequality, (ii) via Cauchy-Schwarz, and (*iii*) and (64) with (65). To finish, we rewrite the first term in (66) like so: recall that if $X_1,\ldots,X_\ell$ are ℓ i.i.d. random variables with the same distribution as some random variable X, then

$$\mathbb{E}[e^X]^\ell = \mathbb{E}[e^{X_1}]\cdots\mathbb{E}[e^{X_\ell}] = \prod_{r=1}^{\ell}\mathbb{E}[e^{X_r}] = \mathbb{E}\left[\prod_{r=1}^{\ell} e^{X_r}\right] = \mathbb{E}\left[e^{\sum_{r=1}^{\ell} X_r}\right],$$

which means that we can say

$$\begin{aligned}
&\left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} Q_{\mathsf{D}}\left(\langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}\right)\right]\right| \\
&\qquad = \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} \sum_{\ell=0}^{\mathsf{D}} b_\ell \langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)}\rangle_{i,k}^{\ell}\right]\right| \\
&\qquad \leq \sum_{\ell=0}^{\mathsf{D}} |b_\ell| \left|\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_0)} \langle e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\rangle_{i,k}^{\ell}\right]\right| \\
&\qquad \leq \sum_{\ell=0}^{\mathsf{D}} |b_\ell| \left|\left\langle\mathbb{E}_{(i,k)}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_r)}\right]\right\rangle_{i,k,\ell}\right| \\
&\qquad \leq \sum_{\ell=0}^{\mathsf{D}} |b_\ell| \sup_{\substack{\beta_1,\ldots,\beta_\ell \\ \in\tilde{\mathcal{S}}}} \left|\mathbb{E}\left[\tilde{U}_i^\intercal \mathcal{D}_i(\beta_0) e^{-\alpha\sum_{r=0}^{\ell}\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta_r)}\right]\right|,
\end{aligned} \tag{67}$$

where the final expectation is unconditional due to independence, and $\langle \cdot \rangle_{i,k,\ell}$ represents the ℓ-dimensional joint expectation with marginals following $\langle \cdot \rangle_{i,k}$. Combining (66) and (67) with the supremum over β_0 in (63), we conclude the result. ■

The following lemma allows us to convert the statement about Gaussian approximation from Assumption 5, which involves k terms, to one that involves arbitrarily many, which will be important when combined with the polynomial derived from the previous lemma.

Lemma 28 (k-to-many Betas) *Let $\ell \geq 1$ and $g : \mathbb{R}^{2\ell k} \to \mathbb{R}$. Then*

1. *If g is bounded Lipschitz and*

$$K \;:=\; \sup_{\substack{f\in\mathcal{F}\\ \theta\in\mathcal{S}^{k-1}}} \; \sup_{\substack{(\beta_1,\ldots,\beta_k)\\ \in\mathcal{S}_p^k}} \left| \mathbb{E} f\left(\sum_{i=1}^{k} \theta_i X_i^\intercal \beta_i\right) - \mathbb{E} f\left(\sum_{i=1}^{k} \theta_i G_i^\intercal \beta_i\right) \right| \;\leq\; 1 \;,$$

   then there exists some $\mathsf{C}_{k,l} > 0$ such that

$$\begin{aligned}&\sup_{\substack{B=(\beta_1,\ldots,\beta_\ell)\\ \in\tilde{\mathcal{S}}^\ell}} \left| \mathbb{E}\left[g(X_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) - g(G_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) \right] \right| \\ &\quad \leq \mathsf{C}_{k,l} \left(\|g\|_{\mathrm{Lip}} + \|g\|_\infty \left(\mathbb{E}[\|X_{\mathcal{B}_i}B\|_\infty + \|G_{\mathcal{B}_i}B\|_\infty + 2\|\tilde{G}_{\mathcal{B}_i}B\|_\infty] + 1\right)\right) K^{\frac{1}{8kl-4}}\end{aligned}$$

   where $\tilde{\mathbf{G}}$ is an independent copy of $\mathbf{G}$ and $\mathcal{F}$ is as in Assumption 5.

2. *If g is locally Lipschitz & square-integrable and Assumption 5 holds, then*

$$\sup_{\substack{B=(\beta_1,\ldots,\beta_\ell)\\ \in\tilde{\mathcal{S}}^\ell}} \left| \mathbb{E}\left[g(X_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) - g(G_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) \right] \right| \xrightarrow{n\to\infty} 0.$$

Proof To prove the first statement, let $g : \mathbb{R}^{2\ell k} \to \mathbb{R}$ be a bounded Lipschitz function. For notational simplicity, we may assume WLOG that the block $\mathcal{B}_i$ begins at index i, meaning

$$X_{\mathcal{B}_i} := \begin{pmatrix} — & X_i & — \\ & \vdots & \\ — & X_{i+k-1} & — \end{pmatrix} \in \mathbb{R}^{k\times p}, \qquad B := \begin{pmatrix} | & & | \\ \beta_1 & \cdots & \beta_\ell \\ | & & | \end{pmatrix} \in \tilde{\mathcal{S}}^\ell \subseteq \mathbb{R}^{p\times\ell}.$$

Let us define the following quantities:

$$\mathcal{M} := B^\intercal \otimes I_{2k} \in \mathbb{R}^{2k\ell\times 2kp} \qquad \boldsymbol{u} := \mathsf{vec}\begin{pmatrix} X_{\mathcal{B}_i} \\ \tilde{G}_{\mathcal{B}_i} \end{pmatrix}, \; \boldsymbol{v} := \mathsf{vec}\begin{pmatrix} G_{\mathcal{B}_i} \\ \tilde{G}_{\mathcal{B}_i} \end{pmatrix} \in \mathbb{R}^{2kp\times 1},$$

where $\mathsf{vec}(A)$ is the vectorized version of a matrix A. This allows us to say that

$$g(X_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) - g(G_{\mathcal{B}_i}B, \tilde{G}_{\mathcal{B}_i}B) = g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{v}).$$

Now, let $\sigma > 0$ and define $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I_{2k\ell})$ independent of all other quantities. Then by the Triangle Inequality,

$$\begin{aligned}\left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{v})\right]\right| \leq \left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{u} + \mathbf{Z})\right]\right| + \left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{v}) - g(\mathcal{M}\boldsymbol{v} + \mathbf{Z})\right]\right| \qquad (68)\\ + \left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - g(\mathcal{M}\boldsymbol{v} + \mathbf{Z})\right]\right|. \qquad (69)\end{aligned}$$

The two quantities in (68) are easily bounded, as

$$\begin{aligned}\left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{u} + \mathbf{Z})\right]\right| &\leq \mathbb{E}|g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{u} + \mathbf{Z})|\\ &\leq \|g\|_{\mathrm{Lip}} \cdot \mathbb{E}\|\mathbf{Z}\|_2\\ &\leq \|g\|_{\mathrm{Lip}} \cdot \sigma\sqrt{2k\ell}, \qquad (70)\end{aligned}$$

and similarly for $\left|\mathbb{E}\left[g(\mathcal{M}\boldsymbol{v}) - g(\mathcal{M}\boldsymbol{v} + \mathbf{Z})\right]\right|$. To control the quantity in (69), for $R > 0$ we consider the functions

$$\begin{aligned}g_R(y) &:= (g(y) - \bar{g}(y))\,\mathbb{I}_{\{\|y\|_\infty \leq R\}}\ ,\\ \bar{g}(y) &:= \textstyle\sum_{r \leq 2k, r' \leq l} \frac{g(y^{(r,r')}(R)) - g(y^{(r,r')}(-R))}{2R} y_{r,r'}\boldsymbol{e}_{r,r'}\ ,\end{aligned}$$

where $y^{(r,r')}(R) \in \mathbb{R}^{2kl}$ is a copy of y with the (r, r')-th coordinate replaced by R, $y_{r,r'}$ is the (r, r')-th coordinate of y and $\boldsymbol{e}_{r,r'} \in \mathbb{R}^{2kl}$ is the standard Euclidean basis vector with 1 at the position (r, r'). The support of the function g_R is contained within the hyperrectangle $\{y \in \mathbb{R}^{2kl} \,|\, \|y\|_\infty \leq R\}$. Moreover, g_R is continuous inside the hyperrectangle and by the choice of $\bar{g}$, g_R agrees on the boundary of the hyperrectangle. This allows us to extend g_R to a continuous and $\|g\|_{\mathrm{Lip}}$-Lipschitz function $\tilde{g}_R$ that is $2R$-periodic in each coordinate. By construction

$$\begin{aligned}&\left|\mathbb{E}[g(\mathcal{M}\boldsymbol{u} + \mathbf{Z})] - \mathbb{E}[\tilde{g}_R(\mathcal{M}\boldsymbol{u} + \mathbf{Z})]\right|\\ &\leq \left|\mathbb{E}[g(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - g(\mathcal{M}\boldsymbol{u} + \mathbf{Z})\mathbb{I}_{\{\|\mathcal{M}\boldsymbol{u}+\mathbf{Z}\|_\infty \leq R\}}]\right| + \left|\mathbb{E}[\bar{g}(\mathcal{M}\boldsymbol{u} + \mathbf{Z})\,\mathbb{I}_{\{\|\mathcal{M}\boldsymbol{u}+\mathbf{Z}\|_\infty \leq R\}}]\right|\\ &\qquad + \left|\mathbb{E}[\tilde{g}_R(\mathcal{M}\boldsymbol{u} + \mathbf{Z})\,\mathbb{I}_{\{\|\mathcal{M}\boldsymbol{u}+\mathbf{Z}\|_\infty > R\}}]\right|\\ &\leq \|g\|_\infty\,\mathbb{P}(\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty > R) + \frac{2kl\,\|g\|_\infty}{R}\mathbb{E}[\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty] + (1 + 2kl)\,\|g\|_\infty\,\mathbb{P}(\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty > R)\\ &\leq \frac{(2 + 4kl)\|g\|_\infty\,\mathbb{E}[\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty]}{R}\ , \qquad (71)\end{aligned}$$

and the same bound holds with $\boldsymbol{u}$ replaced by $\boldsymbol{v}$. We can now approximate $\tilde{g}_R$ coordinate-wise by a truncated Fourier series. Since $\tilde{g}_R$ is continuous and $\|g\|_{\mathrm{Lip}}$-Lipschitz, applying a well-established result on uniform convergence of Fourier series to each coordinate (see e.g. Jackson (1930); Alimov et al. (1992)) says that there exists some absolute constant $C' > 0$ such that, for every $N \in \mathbb{N}$,

$$\|\tilde{g}_R - \tilde{g}_R^{(N)}\|_\infty \ \leq\ C'\|g\|_{\mathrm{Lip}}\,\frac{\log N}{N}\ ,$$

where the truncated Fourier series is given by

$$\tilde{g}_R^{(N)}(y) \ :=\ \textstyle\sum_{w \in \{-N,\ldots,N\}^{2kl}} e^{i\frac{\pi w^\top y}{R}} \left(\frac{1}{(2R)^{2kl}} \int_{[0,2R]^{2kl}} \tilde{g}_R(t) e^{-i\frac{\pi w^\top t}{R}} dt\right).$$

This implies that

$$\begin{aligned}
&\Big|\mathbb{E}[g(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - g(\mathcal{M}\boldsymbol{v} + \mathbf{Z})]\Big| \\
&\leq \ \Big|\mathbb{E}[g(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - \tilde{g}_R(\mathcal{M}\boldsymbol{u} + \mathbf{Z})]\Big| + \Big|\mathbb{E}[g(\mathcal{M}\boldsymbol{v} + \mathbf{Z}) - \tilde{g}_R(\mathcal{M}\boldsymbol{v} + \mathbf{Z})]\Big| \\
&\qquad + \Big|\mathbb{E}[\tilde{g}(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - \tilde{g}_R^{(N)}(\mathcal{M}\boldsymbol{u} + \mathbf{Z})]\Big| + \Big|\mathbb{E}[\tilde{g}(\mathcal{M}\boldsymbol{v} + \mathbf{Z}) - \tilde{g}_R^{(N)}(\mathcal{M}\boldsymbol{v} + \mathbf{Z})]\Big| \\
&\qquad + (\star) \\
&\overset{(71)}{\leq} \ \frac{(2 + 4kl)\,\|g\|_\infty\, \mathbb{E}[\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty + \|\mathcal{M}\boldsymbol{v} + \mathbf{Z}\|_\infty]}{R} + \frac{2C'\|g\|_{\mathrm{Lip}} \log N}{N} + (\star)\,, \qquad (72)
\end{aligned}$$

where

$$\begin{aligned}
(\star) \ &:= \Big|\mathbb{E}[\tilde{g}_R^{(N)}(\mathcal{M}\boldsymbol{u} + \mathbf{Z}) - \tilde{g}_R^{(N)}(\mathcal{M}\boldsymbol{v} + \mathbf{Z})]\Big| \\
&= \Big|\sum_{w\in\{-N,\dots,N\}^{2kl}} \Big(\mathbb{E}[e^{i\frac{\pi w^\top(\mathcal{M}\boldsymbol{u}+\mathbf{Z})}{R}} - e^{i\frac{\pi w^\top(\mathcal{M}\boldsymbol{v}+\mathbf{Z})}{R}}]\Big)\Big(\frac{1}{(2R)^{2kl}}\int_{[0,2R]^{2kl}} \tilde{g}_R(t) e^{-i\frac{\pi w^\top t}{R}} dt\Big)\Big| \\
&= \Big|\sum_{w\in\{-N,\dots,N\}^{2kl}} \Big(\mathbb{E}[e^{i\frac{\pi w^\top(\mathcal{M}\boldsymbol{u})}{R}} - e^{i\frac{\pi w^\top(\mathcal{M}\boldsymbol{v})}{R}}]\Big) e^{-\frac{\pi^2\sigma^2\|w\|^2}{2R^2}} \Big(\frac{1}{(2R)^{2kl}}\int_{[0,2R]^{2kl}} \tilde{g}_R(t) e^{-i\frac{\pi w^\top t}{R}} dt\Big)\Big| \\
&\leq \|g\|_\infty \textstyle\sum_{w\in\{-N,\dots,N\}^{2kl}} \Big|\psi_{\mathcal{M}u}(\frac{\pi w}{R}) - \psi_{\mathcal{M}v}(\frac{\pi w}{R})\Big|\, e^{-\frac{\pi^2\sigma^2\|w\|^2}{2R^2}} \\
&= \|g\|_\infty \textstyle\sum_{w\in\{-\frac{N}{R},\dots,\frac{N}{R}\}^{2kl}} \Big|\psi_{\mathcal{M}u}(\pi w) - \psi_{\mathcal{M}v}(\pi w)\Big|\, e^{-\frac{\pi^2\sigma^2\|w\|^2}{2}} \ . \qquad (73)
\end{aligned}$$

To control the difference of characteristic functions, let $t \in \mathbb{R}^{2k\ell}$, and decompose it as $t = (s, \tilde{s})$ for $s, \tilde{s} \in \mathbb{R}^{k\ell}$. Then we have that

$$\begin{aligned}
|\psi_{\mathcal{M}\boldsymbol{u}}(t) - \psi_{\mathcal{M}\boldsymbol{v}}(t)|^2 &\overset{(i)}{=} \Big(\mathbb{E}\Big(e^{i\mathrm{vec}(X_{\mathcal{B}_i}B)^\top s}\Big)\mathbb{E}\Big(e^{i\mathrm{vec}(\tilde{G}_{\mathcal{B}_i}B)^\top \tilde{s}}\Big) - \mathbb{E}\Big(e^{i\mathrm{vec}(G_{\mathcal{B}_i}B)^\top s}\Big)\mathbb{E}\Big(e^{i\mathrm{vec}(\tilde{G}_{\mathcal{B}_i}B)^\top \tilde{s}}\Big)\Big)^2 \\
&\overset{(ii)}{\leq} 2\Big|\mathbb{E}\Big(e^{i\mathrm{vec}(X_{\mathcal{B}_i}B)^\top s}\Big) - \mathbb{E}\Big(e^{i\mathrm{vec}(G_{\mathcal{B}_i}B)^\top s}\Big)\Big|, \qquad (74)
\end{aligned}$$

where (i) is because the characteristic function factors due to $\tilde{\mathbf{G}} \perp\!\!\!\perp (\mathbf{X}, \mathbf{G})$, and (ii) is because the characteristic function always has modulus in $[0, 1]$, and $(x - y)^2 \leq 2|x - y|$ for $x, y \in [0, 1]$. From here, let us now decompose our vector $s \in \mathbb{R}^{k\ell}$ into k subvectors by defining

$$s_r := s_{r(\ell-1)+1:r\ell} = (s_{r(\ell-1)+1}, \dots, s_{r\ell}) \in \mathbb{R}^\ell$$

for each $r = 1, \dots, k$. Then, we expand

$$
\begin{aligned}
\mathsf{vec}(X_{\mathcal{B}_i} B)^\top s &= \sum_{r=1}^{k} (X_{r+i-1}^\top \beta_1, \dots, X_{r+i-1}^\top \beta_\ell)^\top s_r \\
&= \sum_{r=1}^{k} X_{r+i-1}^\top (B s_r) \\
&= \sum_{r=1}^{k} X_{r+i-1}^\top \sum_{t=1}^{\ell} s_{rt} \beta_t \\
&= \sum_{r=1}^{k} \|s_r\|_1 X_{r+i-1}^\top \sum_{t=1}^{\ell} \frac{|s_{rt}|}{\|s_r\|_1} \mathrm{sgn}(s_{rt}) \beta_t \\
&=: \sum_{r=1}^{k} \|s_r\|_1 X_{r+i-1}^\top \nu_r, \qquad (75)
\end{aligned}
$$

where each $\nu_r := \sum_{t=1}^{\ell} \frac{|s_{rt}|}{\|s_r\|_1} \mathrm{sgn}(s_{rt}) \beta_t$ is in $\mathcal{S}_p$, since if we define

$$
(\beta_{\text{mix}})_{rt} := \mathrm{sgn}(s_{rt}) \beta_t \in \mathcal{S}_p, \quad \lambda_{rt} := \frac{|s_{rt}|}{\|s_r\|_1} \in [0, 1],
$$

then we know that since $\mathcal{S}_p$ is symmetric and convex and $\sum_t \lambda_{rt} = 1$, and since $\mathcal{S}_p$ contains the convex closure of $\tilde{S}$, it must be that

$$
\nu_r = \sum_{t=1}^{\ell} \frac{|s_{rt}|}{\|s_r\|_1} \mathrm{sgn}(s_{rt}) \beta_t = \lambda_{r1} \tilde{\beta}_{r1} + \ldots + \lambda_{r\ell} \tilde{\beta}_{r\ell} \in \mathcal{S}_p.
$$

Thus we conclude that

$$
\begin{aligned}
\sup_B |\psi_{\mathcal{M}\boldsymbol{u}}(t) - \psi_{\mathcal{M}\boldsymbol{v}}(t)|^2 &\overset{(i)}{\leq} 2 \sup_B \left| \mathbb{E}\left(e^{i \sum_r \|s_r\|_1 X_{r+i-1}^\top \beta_r} \right) - \mathbb{E}\left(e^{i \sum_r \|s_r\|_1 G_{r+i-1}^\top \beta_r} \right) \right| \\
&\leq 2 \sup_B \left| \mathbb{E}\left(e^{i \sqrt{\sum_r \|s_r\|_1^2} \sum_r \frac{\|s_r\|_1 X_{r+i-1}^\top \beta_r}{\sqrt{\sum_r \|s_r\|_1^2}}} \right) - \mathbb{E}\left(e^{i \sqrt{\sum_r \|s_r\|_1^2} \sum_r \frac{\|s_r\|_1 G_{r+i-1}^\top \beta_r}{\sqrt{\sum_r \|s_r\|_1^2}}} \right) \right| \\
&\leq 2 \sup_{\theta \in \mathcal{S}^{k-1}} \sup_B \left| \mathbb{E}\left(e^{i \sqrt{\sum_r \|s_r\|_1^2} \sum_r \theta_r X_{r+i-1}^\top \beta_r} \right) - \mathbb{E}\left(e^{i \sqrt{\sum_r \|s_r\|_1^2} \sum_r \theta_r G_{r+i-1}^\top \beta_r} \right) \right| \\
&\overset{(ii)}{\leq} 2 \sqrt{\sum_r \|s_r\|_1^2} \underbrace{\sup_{\substack{f \in \mathcal{F} \\ \theta \in \mathcal{S}^{k-1}}} \sup_{\substack{(\beta_1, \dots, \beta_k) \\ \in \mathcal{S}_p^k}} \left| \mathbb{E} f\left(\sum_{j=1}^{k} \theta_j X_{j+k-1}^\top \beta_j \right) - \mathbb{E} f\left(\sum_{j=1}^{k} \theta_j G_{j+k-1}^\top \beta_j \right) \right|}_{=K}.
\end{aligned}
$$

Here, (i) is via (74) and (75), and (ii) is because the map $x \mapsto c^{-1} e^{icx}$ is 1-Lipschitz and bounded by c^{-1}. Now let $w_{r,r'}$ be the $r(l-1) + r'$-th coordinate of w, where $1 \leq r \leq 2k$ and $1 \leq r' \leq l$. Plugging the above bound into (73), we obtain

$$\begin{aligned}
(\star) \;\; &\leq \|g\|_\infty \;\sqrt{2K}\; \textstyle\sum_{w \in \{-\frac{N}{R},\dots,\frac{N}{R}\}^{2kl}} \Big(\textstyle\sum_{r=1}^{k} \Big(\textstyle\sum_{r'=1}^{l} |\pi w_{r,r'}|\Big)^2\Big)^{1/4} e^{-\frac{\pi^2\sigma^2\|w\|^2}{2}} \\
&\leq \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\; \textstyle\sum_{w \in \{-\frac{N}{R},\dots,\frac{N}{R}\}^{2kl}} e^{-\frac{\pi^2\sigma^2\|w\|^2}{2}} \\
&= \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\Big(1 + 2R \textstyle\sum_{w' \in \{1,\frac{1}{R},\dots,\frac{N}{R}\}} \frac{1}{R}\, e^{-\frac{\pi^2\sigma^2 (w')^2}{2}}\Big)^{2kl} \\
&\leq \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\Big(1 + 2R \int_0^\infty e^{-\frac{\pi^2\sigma^2 (w')^2}{2}}\, dw'\Big)^{2kl} \\
&= \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\Big(1 + \frac{R\sqrt{2}}{\sigma}\Big)^{2kl} .
\end{aligned}$$

Combining this bound with (70) and (72) by the triangle inequality, we get

$$\begin{aligned}
&\Big|\mathbb{E}[g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{v})]\Big| \\
&\leq\; \|g\|_{\mathrm{Lip}}\sigma\sqrt{2kl} + \frac{(2+4kl)\,\|g\|_\infty\, \mathbb{E}[\|\mathcal{M}\boldsymbol{u} + \mathbf{Z}\|_\infty + \|\mathcal{M}\boldsymbol{v} + \mathbf{Z}\|_\infty]}{R} + \frac{2\|g\|_{\mathrm{Lip}} C' \log N}{N} \\
&\qquad + \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\Big(1 + \frac{R\sqrt{2}}{\sigma}\Big)^{2kl} \\
&\leq\; \|g\|_{\mathrm{Lip}}\sigma\sqrt{2kl} + \frac{(2+4kl)\,\|g\|_\infty\, (\mathbb{E}[\|\mathcal{M}\boldsymbol{u}\|_\infty + \|\mathcal{M}\boldsymbol{v}\|_\infty] + 2kl\sigma)}{R} + \frac{2\|g\|_{\mathrm{Lip}} C' \log N}{N} \\
&\qquad + \|g\|_\infty \;\sqrt{2K}\,(kl^2)^{1/4}\, \frac{\pi N}{R}\Big(1 + \frac{R\sqrt{2}}{\sigma}\Big)^{2kl} ,
\end{aligned}$$

where we have used the Markov inequality and a union bound. Now choose

$$\sigma \;=\; K^{\frac{1}{8kl-4}}\,, \qquad N \;=\; R \;=\; \lfloor K^{-\frac{1}{8kl}} \rfloor\,.$$

We get that for some absolute constant $C > 0$,

$$\begin{aligned}
&\Big|\mathbb{E}[g(\mathcal{M}\boldsymbol{u}) - g(\mathcal{M}\boldsymbol{v})]\Big| \\
&\leq\; C\Big(\|g\|_{\mathrm{Lip}}\,(\sqrt{2kl}\; + 1) + \|g\|_\infty(1 + 2kl)(\mathbb{E}[\|\mathcal{M}\boldsymbol{u}\|_\infty + \|\mathcal{M}\boldsymbol{v}\|_\infty] + kl) + \|g\|_\infty\, 3^{2kl}\,(kl^2)^{1/4}\Big)\, K^{\frac{1}{8kl-4}} \\
&\leq\; \mathsf{C}_{k,l}\Big(\|g\|_{\mathrm{Lip}} + \|g\|_\infty\, (\mathbb{E}[\|X_{\mathcal{B}_i} B\|_\infty + \|G_{\mathcal{B}_i} B\|_\infty + 2\|\tilde{G}_{\mathcal{B}_i} B\|_\infty] + 1)\Big)\, K^{\frac{1}{8kl-4}}
\end{aligned}$$

for some $\mathsf{C}_{k,l} > 0$. We also remark that by sub-Gaussianity, the term $\mathbb{E}[\|X_{\mathcal{B}_i} B\|_\infty + \|G_{\mathcal{B}_i} B\|_\infty + 2\|\tilde{G}_{\mathcal{B}_i} B\|_\infty]$ is $O(kl)$.

To prove the second statement of the lemma, we will sketch the outline and refer to Lemma 30 of Montanari and Saeed (2022) for the specific details of a similar approach. We fix $B > 0$, and consider g_B, which forces g to be bounded Lipschitz like so:

$$g_B(\boldsymbol{x}) := g\,(\boldsymbol{x})\,\mathbb{I}\,(\|\boldsymbol{x}\| \leq B) + g\,(B\boldsymbol{x}/\|\boldsymbol{x}\|)\,\mathbb{I}\,(\|\boldsymbol{x}\| > B)\,.$$

We may then apply the first statement of the lemma to g_B, which under Assumption 5 converges to zero. To bound the leftover differences of the form

$$\sup_B \Big|\mathbb{E}\,[g(\boldsymbol{Mu}) - g_B(\boldsymbol{Mu})]\Big|,$$

we use square-integrability of g and the fact that

$$g(\boldsymbol{Mu}) - g_B(\boldsymbol{Mu}) \neq 0 \implies \|\boldsymbol{Mu}\| > B,$$

which by sub-Gaussianity of all $2k\ell$ components of $\boldsymbol{Mu}$ occurs with probability bounded by $\mathsf{C}_1 k\ell \exp\left(-\mathsf{c}_2 B^2/k\ell\right)$. Sending $B \to \infty$ thus concludes the result. ■

Appendix I. Comparing training loss with deleted blocks to the original training loss

Let $M, \tilde{m} \in \mathbb{Z}^+$ be fixed. Define new matrices $\mathbf{X}^M, \mathbf{G}^M \in \mathbb{R}^{n' \times p}$ as

$$\begin{aligned}\mathbf{X}^M &:= (X_1, \dots, X_M, X_{M+\tilde{m}+1}, \dots, X_{2M+\tilde{m}}, X_{2M+2\tilde{m}+1}, \dots)^\top\\ \mathbf{G}^M &:= (G_1, \dots, G_M, G_{M+\tilde{m}+1}, \dots, G_{2M+\tilde{m}}, G_{2M+2\tilde{m}+1}, \dots)^\top.\end{aligned}$$

We show that the training loss $\min_\beta \hat{R}_n(\beta; \mathbf{X})$ and $\min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)$ are close if M is large.

Theorem 29 (Equivalence of losses) *Let $(X_i, y_i(X_i))_{i=1}^n$ and $(G_i, y_i(G_i))_{i=1}^n$ be generated under Assumptions 2-4, where each $G_i \sim \mathcal{N}(\mathbf{0}, \mathrm{Var}(X_i))$. Then if assumption 8(i) is respected then there exists a constant C_d such that*

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right) \leq \mathsf{C}_d \frac{\max(m, \tilde{m})}{M} \sqrt{M + \tilde{m}} \tag{76}$$

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{G}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{G}^M)\right) \leq \mathsf{C}_d \frac{\max(m, \tilde{m})}{M} \sqrt{M + \tilde{m}} \tag{77}$$

If instead assumption 8(ii) is respected then there exists a constant C_d' such that

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right) \leq \mathsf{C}_d' \frac{\max(\mathcal{S}, \tilde{m})}{M} \sqrt{M + \tilde{m}} \tag{78}$$

$$d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{G}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{G}^M)\right) \leq \mathsf{C}_d' \frac{\max(\mathcal{S}, \tilde{m})}{M} \sqrt{M + \tilde{m}} \tag{79}$$

Proof of Theorem 29 We present the proof in the m-dependent case but the proof is identical in the other case. Similarly we show the proof only for $\mathbf{X}$ as the proof is equivalent for $\mathbf{G}$. Let $M \in \mathbb{Z}^+$ be fixed. Define new matrices $\mathbf{X}^M \in \mathbb{R}^{n' \times p}$ as

$$\mathbf{X}^M := (X_1, \dots, X_M, X_{M+\tilde{m}+1}, \dots, X_{2M+\tilde{m}}, X_{2M+2m+1}, \dots)^\top$$

noting that

$$n' \in [n - (r+1)m + 1,\ n - rm] \subset \left[n\frac{M}{M+\tilde{m}} - m,\ n\frac{M}{M+\tilde{m}} + m\right] = [nq - \tilde{m}, nq + \tilde{m}], \tag{80}$$

where $r := \lfloor \frac{n}{M+\tilde{m}} \rfloor$ and $q := \frac{M}{M+\tilde{m}}$. We may also define

$$\mathbf{X}^{\tilde{m}} := (X_{M+1}, \dots, X_{M+\tilde{m}}, X_{2M+\tilde{m}+1}, \dots, X_{2M+2\tilde{m}}, \dots)$$

so that every vector X_i is either in $\mathbf{X}^M$ or $\mathbf{X}^m$. For simplicity we can also write these indexing sets as

$$\begin{aligned} B_M &:= \{1, \ldots, M, M + \tilde{m} + 1, \ldots, 2M + m, \ldots, \} \\ B_m &:= [n] \setminus B_M. \end{aligned}$$

By a triangle inequality argument we note that

$$\begin{aligned} d_{\mathcal{H}}\left(\min_\beta \hat{R}_n(\beta; \mathbf{X}), \min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right) &= \sup_{h \in \mathcal{H}} \left| \mathbb{E}\left[h\left(\min_\beta \hat{R}_n(\beta; \mathbf{X})\right) - h\left(\min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right)\right]\right| \\ &\leq \sup_{h \in \mathcal{H}} \mathbb{E}\left| h\left(\min_\beta \hat{R}_n(\beta; \mathbf{X})\right) - h\left(\min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right)\right| \\ &\overset{(i)}{\leq} \mathbb{E}\left| \min_\beta \hat{R}_n(\beta; \mathbf{X}) - \min_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)\right| \\ &\overset{(ii)}{=:} \mathbb{E}\left| \hat{R}_n(\hat{\beta}; \mathbf{X}) - \hat{R}_{n'}(\tilde{\beta}; \mathbf{X}^M)\right|, \end{aligned} \tag{81}$$

where (i) is via the Lipschitzness of h, and in (ii) we defined the minimizers $\hat{\beta} := \text{argmin}_\beta \hat{R}_n(\beta; \mathbf{X})$ and $\tilde{\beta} := \text{argmin}_\beta \hat{R}_{n'}(\beta; \mathbf{X}^M)$. We first control the difference inside the absolute value of (81) by noting that, by definition of being minimizers,

$$\begin{aligned} \hat{R}_n(\hat{\beta}; \mathbf{X}) - \hat{R}_{n'}(\tilde{\beta}; \mathbf{X}^M) &\leq \left(1 - \frac{n}{n'}\right) \hat{R}_n(\hat{\beta}) + \frac{1}{n'} \sum_{i \in B_m} \ell\left(\eta_i, X_i^\intercal \tilde{\beta}\right) \leq \frac{1}{n'} \sum_{i \in B_m} \ell\left(\eta_i, X_i^\intercal \tilde{\beta}\right) \\ \hat{R}_{n'}(\tilde{\beta}; \mathbf{X}^M) - \hat{R}_n(\hat{\beta}; \mathbf{X}) &\leq \left(1 - \frac{n'}{n}\right) \hat{R}_{n'}(\tilde{\beta}) - \frac{1}{n} \sum_{i \in B_m} \ell\left(\eta_i, X_i^\intercal \hat{\beta}\right) \leq \left(1 - \frac{n'}{n}\right) \hat{R}_{n'}(\tilde{\beta}), \end{aligned}$$

where the terms removed are because of the fact that the risk/loss is always positive and $n > n'$. Now, we will use the fact that for any $x \in \mathbb{R}$ and $a, b > 0$, we have

$$x \leq a, \ -x \leq b \implies |x| \leq a \vee b \leq a + b$$

to say that

$$\left| \hat{R}_n(\hat{\beta}; \mathbf{X}) - \hat{R}_{n'}(\tilde{\beta}; \mathbf{X}^M)\right| \leq \left(1 - \frac{n'}{n}\right) \hat{R}_{n'}(\tilde{\beta}) + \frac{1}{n'} \sum_{i \in B_m} \ell\left(\eta_i, X_i^\intercal \tilde{\beta}\right). \tag{82}$$

For the first term on the right in (82), we note by definition of being a minimizer that

$$\left(1 - \frac{n'}{n}\right) \hat{R}_{n'}(\tilde{\beta}) \leq \left(1 - \frac{n'}{n}\right) \hat{R}_{n'}(0) = \left(1 - \frac{n'}{n}\right) \log(2). \tag{83}$$

For the second term in (82), we have that

$$
\begin{aligned}
\frac{1}{n'}\sum_{i\in B_m}\ell\left(\eta_i, X_i^\intercal\tilde{\beta}\right) &\overset{(i)}{\leq} \frac{1}{n'}\sum_{i\in B_m} 1+|X_i^\intercal\tilde{\beta}| \\
&\leq \frac{n-n'}{n'}+\frac{1}{n'}\sup_{\beta}\sum_{i\in B_m}|X_i^\intercal\beta| \\
&= \frac{n-n'}{n'}+\frac{1}{n'}\sup_{\beta}\sum_{i=1}^{n}|X_i^\intercal\beta|\mathbb{I}(i\in B_m) \\
&\overset{(ii)}{\leq} \frac{n-n'}{n'}+\frac{1}{n'}\sup_{\beta}\sqrt{\sum_{i=1}^{n}|X_i^\intercal\beta|^2}\sqrt{\sum_{i=1}^{n}\mathbb{I}(i\in B_m)} \\
&= \frac{n-n'}{n'}+\frac{\sqrt{n-n'}}{n'}\|\mathbf{X}\|_{\mathcal{S}_p}, \qquad (84)
\end{aligned}
$$

where (i) is because

$$\log(1+e^{-x}) \leq |x|+1 \text{ for all } x\in\mathbb{R},$$

and (ii) is via Cauchy-Schwarz. Combining (83) and (84), we obtain

$$
\begin{aligned}
\mathbb{E}\left|\hat{R}_n(\hat{\beta};\mathbf{X})-\hat{R}_{n'}(\tilde{\beta};\mathbf{X}^M)\right| &\leq \left(1-\frac{n'}{n}\right)\log(2)+\frac{n-n'}{n'}+\frac{\sqrt{n-n'}}{n'}\mathbb{E}\left[\|\mathbf{X}\|_{\mathcal{S}_p}\right] \qquad (85)\\
&\overset{(i)}{\leq} \left(1-\frac{n'}{n}\right)\log(2)+\frac{n-n'}{n'}+\frac{\sqrt{n-n'}}{n'}\sqrt{\mathsf{C}_L mp}, \qquad (86)
\end{aligned}
$$

where (i) is via Jensen's Inequality with Corollary 20. From here we can use (21) to note that, for these quantities involving both n and n', we have

$$n'\in nq\pm\tilde{m} \implies \frac{n'}{n}\in q\pm\frac{\tilde{m}}{n} \implies \lim_{n\to\infty}\frac{n'}{n}=q. \qquad (87)$$

Substituting this into (86) and using that $q=\frac{M}{M+\tilde{m}}$, after some simplification we can conclude that

$$
\begin{aligned}
\limsup_{n\to\infty}\mathbb{E}\left|\hat{R}_n(\hat{\beta};\mathbf{X})-\hat{R}_{n'}(\tilde{\beta};\mathbf{X}^M)\right| &\leq (1-q)\log(2)+\left(\frac{1}{q}-1\right)+\sqrt{\frac{\mathsf{C}_L m\kappa}{q}}\sqrt{\frac{1}{q}-1} \qquad (88)\\
&\leq \mathsf{C}_2\frac{\max(m,\tilde{m})}{M}\sqrt{M+\tilde{m}}. \qquad (89)
\end{aligned}
$$

■

Appendix J. Verifying Assumption 5 for Different Data Augmentation Schemes

J.1. Random Cropping

The data augmentation procedure that we are considering in this subsection is the random cropping method where a portion of the data is randomly set to 0. For a vector $e:=(e_i)\in\{0,1\}^p$ and $x\in\mathbb{R}^p$

we write $e \cdot x := (e_i x_i)$. Let (E_i) be an i.i.d sequence of random vectors in $\{0,1\}^p$ and define the random transformations $\phi_i(x) = E_i \cdot x$. We will prove that Assumption 5 holds for this type of data augmentation procedure under general conditions. Let (Z_i) a sequence of i.i.d vectors satisfying the following condition

$$(H_{\text{cropping}}(\kappa)) \qquad \mathbb{E}(Z_1) = 0, \ \sup_i \|Z_{1,i}\|_4 < \kappa/\sqrt{n}.$$

Define

$$X_i := \phi_i(Z_{\lceil i/k \rceil}).$$

Lemma 30 *Suppose that the assumption $H_{cropping}(\kappa)$ holds and that the entries $(Z_{1,i})$ are locally dependent and write N_i the dependency neighborhood of $Z_{1,i}$. Suppose that $(E_{1,i})$ is locally dependent and write $\tilde{N}_i$ the local dependency neighborhood of $E_{1,i}$. Then if $|N_i| \times |\tilde{N}_i| = o(n^{r/2})$ Assumption 5 holds for a sequence of random Gaussian vector (G_i) with covariance:*

$$cov(G_{i,j}, G_{m,l}) = \begin{cases} p_l(\mathbb{I}(j=l) + \mathbb{I}(j \neq l)p_j)\text{Var}(Z_{1,j}, Z_{1,l}) \text{ if } i = m \\ p_j p_l cov(Z_{1,j}, Z_{1,l}) \text{ if } |i-m| \leq k, \ i \neq m \\ 0 \text{ otherwise} \end{cases}$$

where $p_j = P(E_{1,j} = 1)$.

Proof Note that as the blocks $(X_{mk+1}, \ldots, X_{(m+1)k})$ are identically distributed it is enough to prove Assumption 5 for $m = 0$. Denote $B_{i,j} = \cup_{j \in N_i} \tilde{N}_j \times [|1, k|]$. Then we remark that the sequence (X_i) is locally dependent and that the dependency neighborhood of $X_{i,j}$ is $B_{i,j}$. The desired result follows from Lemma 42 with $q = 1$. ■

J.2. Noise Injection

The data augmentation procedure that we are considering in this subsection is the noise injection method where random Gaussian noise is injected to the entries. For vectors $g := (g_i), x := (x_i) \in \mathbb{R}^p$ we write $g \cdot x := (g_i + x_i)$. Let (n_i) be an i.i.d sequence of random vectors in $\mathbb{R}^p$ such that $n_1 \sim \mathcal{N}(0, \sigma^2/nId)$ and define the random transformations $\phi_i(x) = n_i \cdot x$. We will prove that Assumption 5 holds for this type of data augmentation procedure under general conditions. Let (Z_i) be a sequence of i.i.d vectors satisfying the following condition

$$(H_{\text{noise}}(\kappa)) \ \mathbb{E}(Z_1) = 0, \qquad \sup_i \|Z_{1,i}\|_4 < \kappa/\sqrt{n}.$$

Define $X_i := \phi_i(Z_{\lceil i/k \rceil})$.

Lemma 31 *Suppose that the assumption $H_{noise}(\kappa)$ holds for an absolute constant $\kappa < \infty$ and that the entries $(Z_{1,i})$ are locally dependent. Write N_i the dependency neighborhood of $Z_{1,i}$. Suppose that $(n_{1,i}) \sim N(0, \frac{\sigma^2}{n} Id)$. Then if $|N_i| = o(n^{r/2})$, Assumption 5 holds for a sequence of random Gaussian vector (G_i) with covariance:*

$$cov(G_{i,j}, G_{m,l}) = \begin{cases} cov(Z_{1,j}, Z_{1,l}) + \delta_{j,l}\delta_{i,m}\sigma^2/n \text{ if } i = m \\ cov(Z_{1,j}, Z_{1,l}) \text{ if } |i-m| \leq k, i \neq m \\ 0 \text{ otherwise} \end{cases} .$$

Proof Note that as the blocks $(X_{mk+1}, \dots, X_{(m+1)k})$ are identically distributed it is enough to prove Assumption 5 for $m = 0$. Denote $B_{i,j} = N_i \times [|1, k|]$. Then we remark that the sequence (X_i) is locally dependent and that the dependency neighborhood of $X_{i,j}$ is $B_{i,j}$. The desired result follows from Lemma 42 with $q = 1$. ■

J.3. Random Sign Flipping

The data augmentation procedure that we are considering in this subsection is the random sign-flip method, where a portion of the data is has it sign randomly flipped. For a vector $e := (e_i) \in \{-1, 1\}^p$ and $x \in \mathbb{R}^p$ we write $e \cdot x := (e_i x_i)$. Let (E_i) be an i.i.d sequence of random vectors in $\{-1, 1\}^p$ and define the random transformations $\phi_i(x) = E_i \cdot x$. We will prove that Assumption 5 holds for this type of data augmentation procedure under general conditions. Let (Z_i) a sequence of i.i.d vectors satisfying the following condition

$$(H_{\text{flip}}(\kappa)) \qquad \mathbb{E}(Z_1) = 0, \ \sup_i \|Z_{1,i}\|_4 < \kappa/\sqrt{n}.$$

Define

$$X_i := \phi_i(Z_{\lceil i/k \rceil}).$$

Lemma 32 *Suppose that the assumption $H_{flip}(\kappa)$ holds and that the entries $(Z_{1,i})$ are locally dependent and write N_i the dependency neighborhood of $Z_{1,i}$. Suppose that $(E_{1,i})$ is locally dependent and write $\tilde{N}_i$ the local dependency neighborhood of $E_{1,i}$. Then if $|N_i| \times |\tilde{N}_i| = o(n^{r/2})$ Assumption 5 holds for a sequence of random Gaussian vector (G_i) with covariance:*

$$cov(G_{i,j}, G_{m,l}) = \begin{cases} p^*_{j,l}\mathbb{E}(Z_{1,j}, Z_{1,l}) - (1 - p^*_{j,l})\mathbb{E}(Z_{1,j}, Z_{1,l}) & \text{if i = m} \\ p_{j,l}\mathbb{E}(Z_{1,j}, Z_{1,l}) - (1 - p_{j,l})\mathbb{E}(Z_{1,j}, Z_{1,l})\ if\ |i - m| \le k\ i \neq m \\ 0 \quad otherwise \end{cases}$$

where $p_{j,l} = P(E_{1,j} = 1)^2 + P(E_{1,j} = -1)^2$ and $p^*_{j,l} = P(E_{1,j} = 1, E_{1,l} = 1) + P(E_{1,j} = -1, E_{1,l} = -1)$.

Proof Note that as the blocks $(X_{mk+1}, \dots, X_{(m+1)k})$ are identically distributed it is enough to prove Assumption 5 for $m = 0$. Denote $B_{i,j} = \cup_{j \in N_i} \tilde{N}_j \times [|1, k|]$. Then we remark that the sequence (X_i) is locally dependent and that the dependency neighborhood of $X_{i,j}$ is $B_{i,j}$. The desired result follows from Lemma 42 with $q = 1$. ■

Remark 33 *Note that we do not assume that the probability of having a sign flipped at one position is the same than at any other positions*

J.4. Random Small Permutations

In this section we will show that Assumption 5 holds for a random permutation scheme. In this goal, let (Z_i) be a sequence of centered i.i.d random vectors with independent (not necessarily identically distributed) entries. We assume that the vectors (Z_i) have blocks of identically distributed entries. More precisely we suppose that there is a partition $(B_i)_{i \le M_n}$ of $[|p|]$ in M_n subsets such that the entries $(Z_{1,i})_{i \in B_u}$ are i.i.d for all $u \le M_n$. We choose (π_i) to be an i.i.d sequence of random permutations of $[|n|]$ that preserve the partition, meaning that for all $j, k \leq n$ that do not belong to the same permutation element then $P(\pi_1(j) = k) = 0$. Define

$$X_i := (Z_{\lceil i/k \rceil, \pi_i^{-1}(\ell)}).$$

We will show that this data augmentation scheme satisfies Assumption 5. In this goal we define the following condition

$$(H_{\text{small permutation}}(\kappa)) \qquad \mathbb{E}(Z_1) = 0, \ \sup_i \|Z_{1,i}\|_4 < \kappa/\sqrt{n}.$$

Lemma 34 *Suppose that the assumption $H_{small\, permutation}(\kappa)$ holds. Suppose that* $\max_i |B_i| = o(n^{r/2})$ *Assumption 5 holds for a sequence of random Gaussian vector (G_i) with covariance:*

$$cov(G_{i,j}, G_{m,l}) = \begin{cases} \sum_{k \in B_{B^{-1}(j)}} p_k^2 \text{var}(\text{Z}_{1,\text{k}}) \ if \, |\text{i} - \text{m}| \leq \text{k} \\ 0 \ otherwise \end{cases}.$$

where $B^{-1}(j)$ denotes the unique index such that $j \in B_{B^{-1}(j)}$ and where $p_k := P(\pi(j) = k)$

Proof Note that as the blocks $(X_{mk+1}, \ldots, X_{(m+1)k})$ are identically distributed it is enough to prove Assumption 5 for $m = 0$. Denote $B_{i,j} = B_\ell \times [|1, k|]$ if $i \in P_\ell$. Then we remark that the sequence (X_i) is locally dependent and that the dependency neighborhood of $X_{i,j}$ is $B_{i,j}$. The desired result follows from Lemma 42 with $q = 1$. ■

J.5. Random Large Permutations

In this section we will show that Assumption 5 holds for a random permutation scheme. In this goal, let (Z_i) be a sequence of centered i.i.d random vectors with independent (not necessarily identically distributed) entries. We assume that the vectors (Z_i) have blocks of identically distributed entries. More precisely we suppose that there is a partition $(B_i)_{i \leq M}$ of $[|p|]$ in M subsets such that the entries $(Z_{1,i})_{i \in B_u}$ are i.i.d for all $u \leq M$. Contrary to the previous section we will have $M << n$. We choose (π_i) to be an i.i.d sequence of random permutations of $[|n|]$ that preserve the partition, meaning that for all $j, k \leq n$ that do not belong to the same permutation element then $P(\pi_1(j) = k) = 0$. Moreover we assume that for all $j, k \in B_u$ in the same partition we have $P(\pi_1(j) = k) = 1/|B_u|$. Define

$$X_i := (Z_{\lceil i/k \rceil, \pi_i^{-1}(\ell)}).$$

Lemma 35 *Suppose that there exists an absolute constant $\kappa < \infty$ such that* $\max_i \|Z_{1,i}\|_3 \leq \kappa/\sqrt{n}$ *and* $\max |B_u| / \min |B_u| < \infty$ *Assumption 5 holds for a sequence of random Gaussian vectors that are such that*

$$cov(G_{i,j}, G_{m,l}) = \sum_u \frac{\sigma_u^2 \mathbb{I}(|i - m| \leq k)}{|B_u|^2}$$

where $\sigma_u^2 = \text{Var}(Y_{1,l})$ *where $l \leq p$ is chosen to be in $l \in B_u$.*

Proof Note that as the blocks $(X_{mk+1}, \dots, X_{(m+1)k})$ are identically distributed it is enough to prove Assumption 5 for $m = 0$. For all $(\theta_i) \in \mathcal{S}^{k-1}$ and all $(\beta_i) \in \mathcal{S}_p^k$ we have

$$\sum_{i\le k} \theta_i X_i^T \beta_i = \sum_{i\le k} \theta_i \sum_l X_{i,l}\beta_{i,l} = \sum_l Z_{1,l} \sum_{i\le k} \theta_i \beta_{i,\pi_i(l)}.$$

Now we notice that conditionally on (π_i) the random variables $(Z_{1,l} \sum_{i\le k} \theta_i\beta_{i,\pi_i(l)})$ are independent. Moreover we observe that

$$\mathbb{E}(Z_{1,l} \sum_{i\le k} \theta_i\beta_{i,\pi_i(l)} \big| (\pi_i)) = 0$$

and for all $l, m \le p$ we have

$$\begin{aligned} \mathrm{cov}(Z_{1,l} \sum_{i\le k} \theta_i\beta_{i,\pi_i(l)}, Z_{1,m} \sum_{i\le k} \theta_i\beta_{i,\pi_i(l)} \big| \pi) &= \delta_{l,m}(\sum_{i\le k} \theta_i\beta_{i,\pi_i(l)})^2 \mathrm{Var}(Z_{1,l}) \\ \mathbb{E}(\big|Z_{1,l} \sum_{i\le k} \theta_i\beta_{i,\pi_i(l)}\big|^3 \big| \pi) &\le \sqrt{k}^3 \max_i \|Z_{1,i}\|_3^3 \sqrt{\sum_{i\le k} \beta^2_{i\pi_i(l)}}^3 \\ &\le k^4 L p^{3/2-3r/2} \max_i \|Z_{1,i}\|_3^3 \end{aligned}$$

where for the last inequality we used Assumption 4. Define

$$\hat\sigma^2 := \sum_{\ell\le p} \mathrm{Var}(Z_{1,\ell})(\sum_{i\le k} \theta_i\beta_{i,\pi_i(\ell)})^2 = \sum_u \sigma_u^2 \sum_{\ell\in B_u} (\sum_{i\le k} \theta_i\beta_{i,\pi_i(\ell)})^2.$$

According to Lemma 41 we have that there exists an absolute constant $\kappa > 0$ such that

$$\sup_{f\in\mathcal{F}} \Big|\mathbb{E}\Big(f(\sum_{i\le k} \theta_i X_i^T \beta_i)\Big|(\pi_i)\Big) - \mathbb{E}_{Z\sim\mathcal{N}(0,\hat\sigma^2)}\Big(f(Z)|(\pi_i)\Big)\Big| \le \kappa^{1/3} p^{1/2-r/2} \max_i \|Z_{1,i}\|_3. \tag{90}$$

Now using the definition of $\hat\sigma^2$ we observe that

$$\mathbb{E}(\hat\sigma^2) = \sum_u \sigma_u^2 \frac{1}{|B_u|^k} \sum_{\ell_1,\dots,\ell_k\in B_u} (\sum_{i\le k} \theta_i\beta_{i,\ell_i})^2 \tag{91}$$

To show that $\hat\sigma^2$ converges to $\mathbb{E}(\hat\sigma^2)$ we will proceed by showing that for all $S \le k$ we have $\mathbb{E}(\hat\sigma^2|\pi_1, \dots, \pi_S)$ concentrates around $\mathbb{E}(\hat\sigma^2|\pi_1, \dots, \pi_{S-1})$. The desired outcome then results from a telescopic sum argument. In this goal we first notice that

$$\begin{aligned} &\mathbb{E}(\hat\sigma^2|\pi_1, \dots, \pi_S) - \mathbb{E}(\hat\sigma^2|\pi_1, \dots, \pi_{S-1}) \\ &= \sum_u \sigma_u^2 \sum_{\ell\in B_u} \sum_{i\ne S} \theta_i\theta_S \mathbb{E}\Big(\beta_{i,\pi_i(\ell)}\beta_{S,\pi_S(\ell)}|\pi_1, \dots, \pi_S\Big) - \mathbb{E}\Big(\beta_{i,\pi_i(\ell)}\beta_{S,\pi_S(\ell)}|\pi_1, \dots, \pi_{S-1}\Big) \end{aligned}$$

To bound this set $I \sim \mathrm{unif}([|n|])$ and $J \sim \mathrm{unif}(B_{B^{-1}(I)})$ where we denote $B^{-1}(i)$ the index u such that $i \in B_u$. Define $\pi'_S = \pi_S \circ (I, J)$ and $\pi'_i = \pi_i$ for all $i \ne S$. Define

$$(\hat\sigma')^2 := \sum_{\ell\le p} \mathrm{Var}(Y_{1,l})(\sum_{i\le k} \theta_i\beta_{i,\pi'_i(\ell)})^2$$

then we notice that $(\hat{\sigma}, \hat{\sigma}')$ forms an exchangeable pair and that

$$\begin{aligned}
&\mathbb{E}(\hat{\sigma}^2|\pi_{1:S}) - \mathbb{E}((\hat{\sigma}')^2|\pi_{1:S}) \\
&= \sum_u \sigma_u^4 \sum_{\ell \in B_u} \sum_{i \neq S} \theta_i \theta_S \, \mathbb{E}\Big(\beta_{i,\pi_i(\ell)}\beta_{S,\pi_S(\ell)}|\pi_1, \ldots, \pi_S\Big) - \mathbb{E}\Big(\beta_{i,\pi_i(\ell)}\beta_{S,\pi_S(\ell)}|\pi_1, \ldots, \pi_{S-1}\Big).
\end{aligned}$$

Moreover we observe that

$$\begin{aligned}
\hat{\sigma}^2 - (\hat{\sigma}')^2 \leq \sigma^2_{B^{-1}(I)} &\sum_{i \neq S} \theta_i \theta_S \left(\beta_{i,\pi_i(I)}\beta_{S,\pi_S(I)} - \beta_{i,\pi_i(I)}\beta_{S,\pi_S(J)}\right) \\
&+ \sigma^2_{B^{-1}(I)} \sum_{i \neq S} \theta_i \theta_S \left(\beta_{i,\pi_i(J)}\beta_{S,\pi_S(J)} - \beta_{i,\pi_i(J)}\beta_{S,\pi_S(I)}\right)
\end{aligned}$$

Hence using Chatterjee (2005) we obtain that

$$\begin{aligned}
\mathrm{Var}\Big(\mathbb{E}(\hat{\sigma}^2\Big|\pi_1, \ldots, \pi_S)\Big|\pi_1, \ldots, \pi_{S-1}\Big) &\leq \mathbb{E}\Big([\hat{\sigma}^2 - (\hat{\sigma}')^2]^2\Big) \\
&\leq 4 \frac{\max_u \sigma_u^2}{p \min_u |B_u|} \sum_{\ell \leq p} \sum_{k \in B_{B^{-1}(\ell)}} \Big(\sum_{i \leq k} \theta_i \theta_S \beta^2_{i,\pi_i(\ell)}(\beta_{S,\pi_S(\ell)} - \beta_{S,\pi_S(k)})\Big)^2 \\
&\leq 8k \frac{\max_u \sigma_u^4}{p \min_u |B_u|} \sum_{\ell} \sum_{k \in B_{B^{-1}(\ell)}} \theta_S^2 \sum_{i \leq k} \theta_i^2 \beta^2_{i,\pi_i(\ell)}(\beta^2_{S,\pi_S(\ell)} + \beta^2_{S,\pi_S(k)}) \\
&\overset{(a)}{\leq} 8k \frac{\max_u \sigma_u^4 L}{p^r \min_u |B_u|} \sum_{\ell} \sum_{k \in B_{B^{-1}(\ell)}} (\beta^2_{S,\pi_S(\ell)} + \beta^2_{S,\pi_S(k)}) \\
&\leq 16k \frac{\max_u \sigma_u^4 L \max_u |B_u|}{p^r \min_u |B_u|} \sum_{\ell} \beta^2_{S,\ell} \\
&\leq 16k \frac{\max_u \sigma_u^4 S^2 \max_u |B_u| p^{1-r}}{\min_u |B_u|}
\end{aligned}$$

where (a) is a result of Assumption 4. Hence as every function $f \in \mathcal{F}$ is Lipschitz we obtain that

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}\Big(f\big(\sum_{i \leq k} \theta_i X_i^T \beta_i\big)\Big) - \mathbb{E}_{Z \sim \mathcal{N}(0,\sigma^2)}\Big(f(Z)\Big) \right| \leq \kappa p^{1/2 - r/2} \max_i \|Z_{1,i}\|_3 + 4\sqrt{k \frac{\max_u \sigma_u^4 L^2 \max_u |B_u| p^{1-r}}{\min_u |B_u|}}$$

Finally we note that $\sum_{i \leq k} \theta_i G_i^T \beta_i \sim N(0, \sigma^2)$ and the required result is hence deduced. ■

Appendix K. Auxiliary Lemmas

K.1. m-dependent Bernstein

The first lemma aims to extend the classic Bernstein's Inequality from the usual independent setting to m-dependence, which of course also then includes block dependence.

Lemma 36 (m-dependent Bernstein) *Let $Z_1, \dots, Z_n$ be centered, sub-exponential, m-dependent random variables, and $\mathsf{K} := \sup_i \|Z_i\|_{\psi_1}$. Then there exists a universal constant $c > 0$ such that for $t \geq 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq t\right) \leq 4 \cdot \exp\left[-\frac{cn}{m}\left(\frac{t}{\mathsf{K}} \wedge \frac{t^2}{\mathsf{K}^2}\right)\right].$$

Proof Define $r := \lfloor n/m \rfloor$, and assume without loss of generality that r is even. For each $j = 1, \dots, r$, let

$$Y_j := \sum_{\ell=(j-1)m+1}^{jm} Z_\ell.$$

Also, define $Y_{r+1} := \sum_{\ell=rm+1}^{n} Z_\ell$. By construction, the set $(Y_1, Y_3, \dots, Y_{r+1})$ is independent, as is the set $(Y_2, Y_4, \dots, Y_r)$. We can then see that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i\right| \geq t\right) &= \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{r+1} Y_j\right| \geq t\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=0}^{r/2} Y_{2j+1}\right| \geq \frac{t}{2}\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^{r/2} Y_{2j}\right| \geq \frac{t}{2}\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{\frac{r}{2}+1}\sum_{j=0}^{r/2} Y_{2j+1}\right| \geq \frac{nt}{r+2}\right) + \mathbb{P}\left(\left|\frac{1}{\frac{r}{2}}\sum_{j=1}^{r/2} Y_{2j}\right| \geq \frac{nt}{r}\right) \qquad (92)
\end{aligned}$$

By the Triangle Inequality, we know that each Y_j is still still sub-exponential with norm

$$\|Y_j\|_{\psi_1} \leq \sum_{\ell=(j-1)m+1}^{jm} \|Z_\ell\|_{\psi_1} \leq m\mathsf{K}.$$

To bound the first term in (92), we use Corollary 2.8.3 of Vershynin (2018) to say that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{1}{\frac{r}{2}+1}\sum_{j=0}^{r/2} Y_{2j+1}\right| \geq \frac{nt}{r+2}\right) &\leq 2 \cdot \exp\left[-c\left(\frac{r}{2}+1\right)\left(\frac{nt}{m(r+2)\mathsf{K}} \wedge \frac{n^2t^2}{m^2(r+2)^2\mathsf{K}^2}\right)\right] \\
&\leq 2 \cdot \exp\left[-\frac{cn}{m}\left(\frac{t}{2\mathsf{K}} \wedge \frac{t^2}{4\mathsf{K}^2}\right)\right],
\end{aligned}$$

where above we used the fact that $r + 2 \geq \frac{n}{m}$. We conclude by noting that this inequality also holds for the second term in (92) by the same reasoning. ■

K.2. Lemmas for mixing processes

In this subsection, we present some useful lemmas to deal with β-mixing processes. The first one allows one to relate the expectation of a bounded function of mixing data to the one of independent blocks.

Lemma 37 (Yu (1994) (Corollary 2.7)) *Let $K \geq 1$ be an integer and let $r_1 \leq s_1 \leq r_2 \leq \cdots \leq s_K$ be a sequence of integers. Let $\tilde{X} = (\tilde{X}_i)$ be a stochastic process taking value in $\mathbb{R}^p$. Suppose that $h : \prod_{\ell \leq K} \mathbb{R}^{p|s_\ell - r_\ell|} \to \mathbb{R}$ is measurable function. Let Q be the distribution of $(\tilde{X}_j)_{j \in \cup_\ell [r_\ell, s_\ell]}$ and Q_i be the marginal distribution of $(\tilde{X}_j)_{j \in [r_i, s_i]}$. Let $\beta_{\max} = \sup_{1 \leq i \leq K-1} \beta(k_i)$, where $k_i = r_{i+1} - s_i$, and the P the product measure $P = \prod_{i=1}^K Q_i$. Then if $\|h\|_\infty$ we have,*

$$|\mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{Y \sim P}[h(Y)]| \leq (K-1)B\beta_{\max}$$

We adapt this lemma to hold for functions with bounded moments.

Lemma 38 *Let $K \geq 1$ be an integer and let $r_1 \leq s_1 \leq r_2 \leq \cdots \leq s_K$ be a sequence of integers. Let $\tilde{X} = (\tilde{X}_i)$ be a stochastic process taking value in $\mathbb{R}^p$. Suppose that $h : \prod_{\ell \leq K} \mathbb{R}^{p|s_\ell - r_\ell|} \to \mathbb{R}$ is measurable function. Let Q be the distribution of $(\tilde{X}_j)_{j \in \cup_\ell [r_\ell, s_\ell]}$ and Q_i be the marginal distribution of $(\tilde{X}_j)_{j \in [r_i, s_i]}$. Let $\beta_{\max} = \sup_{1 \leq i \leq K-1} \beta(k_i)$, where $k_i = r_{i+1} - s_i$, and the P the product measure $P = \prod_{i=1}^K Q_i$. Then, if there exists $L > 1$ such that $|\mathbb{E}_{X \sim Q}[h(X)^L], \mathbb{E}_{Y \sim P}[h(Y)^L] < \infty$ we have*

$$|\mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{Y \sim P}[h(Y)]| \leq 2((K-1)\beta_{\max})^{1-1/L}(\mathbb{E}_{X \sim Q}(h(X)^L))^{1/L} + (\mathbb{E}_{Y \sim P}(h(Y)^2))^{1/L}.$$

Proof Let $C > 0$ be a positive constant then define $\bar{h}^C$ as the following function $\bar{h}^C(x) = h(x)\mathbb{I}(|h(x)| \leq C)$. Now according to Lemma 37 we have that

$$|\mathbb{E}_{X \sim Q}[\bar{h}^C[X]] - \mathbb{E}_Y \sim P[\bar{h}^C[Y]]| \leq (K-1)C\beta_{\max}$$

Now, moreover, note that

$$|\mathbb{E}_{X \sim Q}\Big(h(X)\mathbb{I}(|h(X)| > C)\Big) - \mathbb{E}_{Y \sim P}\Big(h(Y)\mathbb{I}(|h(Y)| > C)\Big)| \leq \frac{1}{C^{L-1}}\Big\{\mathbb{E}_{X \sim Q}(h(X)^L) + \mathbb{E}_{Y \sim P}(h(Y)^L)\Big\}$$

Hence, by triangle inequality, we obtain that

$$|\mathbb{E}_{X \sim Q}[h(X)] - \mathbb{E}_{Y \sim P}[h(Y)]| \leq (K-1)C\beta_{\max} + \frac{1}{C^{L-1}}\Big\{\mathbb{E}_{X \sim Q}(h(X)^L) + \mathbb{E}_{Y \sim P}(h(Y)^L)\Big\}.$$

By choosing $C := \Big(\frac{(L-1)\Big\{\mathbb{E}_{X \sim Q}(h(X)^L) + \mathbb{E}_{Y \sim P}(h(Y)^L)\Big\}}{(K-1)\beta_{\max}}\Big)^{1/L}$ we get the desired result. ■

The next lemma establishes concentration for some function of $\beta-$mixing random variables.

Lemma 39 *Let $(\mathbf{X}_i)$ be a sequence of random variables with mixing coefficients $(\beta(i))$. Assume that there exists a process $(\mathbf{Z}_i)$ of centered independent random vectors such for every $i \in \mathbb{N}$ there exists constants $(c_{i,j})$ such that we have*

$$\mathbf{X}_i = \sum_{j \in \mathbb{N}} c_{i,j}\mathbf{Z}_j.$$

Suppose that $K := \sup_j \sqrt{p}\|\mathbf{Z}_j\|_{\psi_2} < \infty$ and that $\mathrm{Var}(\mathrm{Z_j}) = \Sigma/\mathrm{p}$ is such that $\lambda_{\min}(\Sigma) \geq \underline{c} > 0$. Moreover assume that there exists $\epsilon > 0$ such that $\mathcal{S} := \sum_{\ell < \infty} \beta(\ell)^{\frac{\epsilon}{2+\epsilon}} < \infty$. Then there exists constants $C, \tilde{C}, \tilde{C}'$ such that for all $\beta \in \mathcal{S}_p$

$$P(\Big|\sum_{i \leq n}(\mathbf{X}_i^T\beta)^2 - \mathbb{E}((\mathbf{X}_i^T\beta)^2)\Big| \geq t) \leq 2e^{-\frac{Ct\underline{c}}{K\mathcal{S}\|\beta\|_2^2/p}\min\Big(\frac{t\underline{c}}{\tilde{C}Kn\|\beta\|^2/p}, \frac{1}{\tilde{C}'}\Big)}.$$

Let $(\mathbf{G}_i)$ *be a sequence of centered Gaussian random variables such that for all* $t > 0$, $\mathrm{Cov}(\mathbf{G}) = \mathrm{Cov}(\mathbf{X})$ *then we obtain that there exists universal constants* $C_2, \tilde{C}_2, \tilde{C}_2' > \infty$ *such that for all* $\beta \in \mathcal{S}_p$

$$P(\Big|\sum_{i\leq n}(\mathbf{G}_i^T\beta)^2 - \mathbb{E}((\mathbf{G}_i^T\beta)^2)\Big| \geq t) \leq 2e^{-\frac{C_2 t}{S\|\beta\|_2^2/p}\min\left(\frac{t\tilde{C}_2}{n\|\beta\|_2^2/p},\tilde{C}_2'\right)}.$$

Proof Firstly we remark that we can assume without loss of generality that there exists an $M \in \mathbb{N}$ such that $\mathbf{X}_i = \sum_{j\leq M} c_{i,j}Z_j$ for all $i \leq n$. We define $F_m := \sum_\ell \beta_\ell Z_{m,\ell}$ and remark that the process (F_m) is still a sequence of independent random variables that are $K\|\beta\|^2/p$ subGaussian. Moreover we remark that for all $i \leq n$ we have

$$\beta^T\mathbf{X}_i = \sum_{m\leq M} c_{i,m}F_m.$$

Define $C_M := (c_{i,j})_{i\leq n, j\leq M}$ then we immediately note that

$$\sum_{i\leq n}(\beta^T\mathbf{X}_i)^2 = F_m^T(C_M^TC_M)F_m.$$

Hence using Theorem 1.1 of Rudelson and Vershynin (2013) we remark that there exists $C > 0$ such that

$$P\Big(\Big|\sum_{i\leq n}(\mathbf{X}_i^T\beta)^2 - \mathbb{E}((\mathbf{X}_i^T\beta)^2)\Big| \geq t\Big) \leq 2e^{-\frac{tC}{K\|\beta\|^2/p}\min\left(\frac{t}{K\|\beta\|^2/p\|C_M^TC_M\|_{HS}^2},\frac{1}{\|C_M^TC_M\|}\right)}.$$

We note that

$$(C_M^TC_M)_{m_1,m_2} = \sum_{i\leq n} c_{i,m_1}c_{j,m_2}.$$

Hence we remark that

$$\begin{aligned}\|C_M^TC_M\|_{HS}^2 &= \sum_{m_1,m_2\leq M}(\sum_i c_{i,m_1}c_{i,m_2})^2\\ &= \sum_{i,j\leq n}(\sum_m c_{i,m}c_{j,m})^2\end{aligned}$$

We observe that if we define $\nu := (1,\dots,1)^T$ then we have $\mathrm{cov}(\mathbf{X}_{\mathrm{i}}^{\mathrm{T}}\nu, \mathbf{X}_{\mathrm{j}}^{\mathrm{T}}\nu) = \sum_{\mathrm{m}} \mathrm{c}_{\mathrm{i,m}}\mathrm{c}_{\mathrm{j,m}}\nu^{\mathrm{T}}\Sigma\nu$. As we assumed $\lambda_{min}(\Sigma) > \underline{c}$ we obtain that $\nu^T\Sigma\nu \geq \underline{c}$. This implies that

$$\|C_M^TC_M\|_{HS}^2 \leq (\underline{c})^{-2}\sum_{i,j\leq n}\mathrm{Cov}(\mathbf{X}_{\mathrm{i}}^{\mathrm{T}}\nu, \mathbf{X}_{\mathrm{j}}^{\mathrm{T}}\nu)^2.$$

To further bound this, we note that according to Lemma 26 of Austern and Orbanz (2022) we know that

$$|\mathrm{cov}(\nu^{\mathrm{T}}\mathbf{X}_{\mathrm{i}}, \nu^{\mathrm{T}}\mathbf{X}_{\mathrm{j}})| \leq 4\beta(\mathrm{i}-\mathrm{j})^{\frac{\epsilon}{2+\epsilon}}\max_{\mathrm{i}}\|\nu^{\mathrm{T}}\mathbf{X}_{\mathrm{i}}\|_{2+\epsilon}^2,$$

which directly implies, coupled with the fact that $\mathbf{X}_i^T\nu$ is a $K_{\mathbf{X}}$ subgaussian random variable, that there exists a constant $\tilde{C} > 0$ such that

$$\begin{aligned}\|C_M^TC_M\|_{HS}^2 &\leq 16(\underline{c})^{-2}\sum_{i,j\leq n}\mathrm{Cov}(\mathbf{X}_{\mathrm{i}}^{\mathrm{T}}\nu, \mathbf{X}_{\mathrm{j}}^{\mathrm{T}}\nu)^2\beta(\mathrm{i}-\mathrm{j})^{\frac{2\epsilon}{2+\epsilon}}\max_{\mathrm{i}}\|\nu^{\mathrm{T}}\mathbf{X}_{\mathrm{i}}\|_{2+\epsilon}^4\\ &\leq \tilde{C}n(\underline{c})^{-2}\mathcal{S}\end{aligned}$$

Similarly we obtain that there exists a constant $\tilde{C}'$ such that

$$\|C_m^T C_M\| \leq \sup_{m_1} \sum_{m_2 \leq M} c_{i,m_1}, c_{i,m_2} \leq \tilde{C}'(\underline{c})^{-1}\mathcal{S}.$$

Combining this together we obtain that

$$P\Big(\big|\sum_{i\leq n}(\mathbf{X}_i^T\beta)^2 - \mathbb{E}((\mathbf{X}_i^T\beta)^2)\big| \geq t\Big) \leq 2e^{-\frac{Ct\underline{c}}{K\mathcal{S}\|\beta\|_2^2/p}\min\left(\frac{t\underline{c}}{\tilde{C}Kn\|\beta\|^2/p},\frac{1}{\tilde{C}'}\right)}.$$

To prove the second statement, we observe that $(\mathbf{G}_i^T\beta)$ is still a Gaussian process. Hence if we let $\mathbf{N} := (N_i)_{i\leq n}$ be a vector of standard normal we obtain that $\sum_{i\leq n}(\mathbf{G}_i^T\beta)^2 \overset{d}{=} \mathbf{N}^T\Sigma_n\mathbf{N}$ where $\Sigma_n := \mathrm{Cov}((\mathbf{G}_\mathrm{i}^\mathrm{T}\beta)_{\mathrm{i}=1}^{\mathrm{n}})$. Hence for every $t \geq 0$ we have that

$$P(\big|\sum_{i\leq n}(\mathbf{G}_i^T\beta)^2 - \mathbb{E}((\mathbf{G}_i^T\beta)^2)\big| \geq t) \leq P(\Big|\mathbf{N}^T\Sigma_n\mathbf{N} - \mathbb{E}(\mathbf{N}^T\Sigma_n\mathbf{N})\Big| \geq t)$$

To further bound this the first step is to bound. In this goal we remark that $(\Sigma_n)_{i,j} = \mathrm{Cov}(\mathbf{G}_\mathrm{i}^\mathrm{T}\beta, \mathbf{G}_\mathrm{j}^\mathrm{T}\beta) = \mathrm{Cov}(\mathbf{X}_\mathrm{i}^\mathrm{T}\beta, \mathbf{X}_\mathrm{j}^\mathrm{T}\beta)$. Now once again using Lemma 26 of Austern and Orbanz (2022) we know that

$$|\mathrm{cov}(\beta^\mathrm{T}\mathbf{X}_\mathrm{i}, \beta^\mathrm{T}\mathbf{X}_\mathrm{j})| \leq 4\beta(\mathrm{i}-\mathrm{j})^{\frac{\epsilon}{2+\epsilon}} \max_\mathrm{i} \|\beta^\mathrm{T}\mathbf{X}_\mathrm{i}\|_{2+\epsilon}^2 \leq 4\|\beta\|_2^2/\mathrm{p}\beta(\mathrm{i}-\mathrm{j})^{\frac{\epsilon}{2+\epsilon}} \max_{\mathrm{i},\|\mu\|\leq\sqrt{p}} \|\nu\mathbf{X}_\mathrm{i}\|_{2+\epsilon}^2.$$

This directly implies that there exists another constant $\tilde{C}_2$ such that

$$\|\Sigma_n\|_F^2 \leq (\|\beta\|_2^2/p)^2\tilde{C}_2\sqrt{n}\mathcal{S}.$$

We moreover similarly notice that there exists another constant $\tilde{C}_2'$ such that

$$\|\Sigma_n\| \leq \max_i \sum_{j\leq n} \mathrm{cov}(\mathbf{X}_\mathrm{i}^\mathrm{T}\beta, \mathbf{X}_\mathrm{j}^\mathrm{T}\beta) \leq \|\beta\|_2^2/\mathrm{p}\mathcal{S}\tilde{\mathrm{C}}_2'.$$

Hence using the Hanson-Wright Inequality (Theorem 1.1 of Rudelson and Vershynin (2013)), we obtain that there exists a constant $C_2 > 0$ such that

$$P(\Big|\mathbf{N}^T\Sigma_n\mathbf{N} - \mathbb{E}(\mathbf{N}^T\Sigma_n\mathbf{N})\Big| \geq t) \leq 2e^{-\frac{C_2 t}{S\|\beta\|_2^2/p}\min\left(\frac{t\tilde{C}_2}{n\|\beta\|_2^2/p},\tilde{C}_2'\right)}.$$

■

K.3. Smoothing lemma

Lemma 40 *For every $\delta > 0$, there exists a family of continuous and $\frac{1}{\delta}$-Lipschitz $\mathbb{R} \to \mathbb{R}$ functions $(h_{\tau;\delta})_{\tau\in\mathbb{R}}$ such that, for any $x, y \in \mathbb{R}$,*

(i) $h_{\tau;\delta}(x) \leq \mathbb{I}\{x < \tau\} \leq h_{\tau+\delta;\delta}(x)$;

(ii) $\Big|h_{\tau;\delta}(x) - \mathbb{I}\{x < \tau\}\Big| \leq \mathbb{I}\{x \in [\tau - \delta, \tau)\}$.

Proof of Lemma 40. We construct $h_{\tau;\delta}$ as

$$h_{\tau;\delta}(x) \;:=\; \begin{cases} 1 & \text{if } x < \tau - \delta \,, \\ \frac{\tau - x}{\delta} & \text{if } x \in [\tau - \delta, \tau) \,, \\ 0 & \text{if } x \geq \tau \,, \end{cases}$$

which satisfies (i) automatically. To prove (ii), we first note that

$$\mathbb{I}\{x < \tau - \delta\} \;\leq\; h_{\tau;\delta}(x) \;\leq\; \mathbb{I}\{x < \tau\} \,,$$

which implies the desired bound that

$$\Big| h_{\tau;\delta}(x) - \mathbb{I}\{x < \tau\} \Big| \leq \mathbb{I}\{x \in [\tau - \delta, \tau)\} \,.$$

■

K.4. Additional Lindeberg Lemma

Lemma 41 *Suppose that (X_i) are independent, centered random variables. We obtain that there is an absolute constant $\kappa > 0$ such that*

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E} f\left(\sum_{j \leq n} X_j\right) - \mathbb{E} f\left(\sum_{j \leq n} Z_j\right) \right| \tag{93}$$

$$\leq \left(\frac{\kappa n}{\epsilon^2}\right)^{1/3} \max_{j \leq n} \|X_j\|_3, \tag{94}$$

where (Z_i) is an independent sequence of independent and centered Gaussian random variables.

Proof Let $f \in \mathcal{F}$, and let $\epsilon > 0$. Define $f_\epsilon(u) = \frac{1}{4\epsilon^2} \int_{u-\epsilon}^{u+\epsilon} \int_{t-\epsilon}^{t+\epsilon} f(y) dy dt$. We remark that f_ϵ is three times differentiable and as f is Lipschitz we have $\sup_x |f_\epsilon(x) - f(x)| \leq 2\epsilon$.

Write

$$X_j(t) := \sqrt{t} X_j + \sqrt{1 - t} G_j$$

and

$$X_m^{j,0}(t) := \begin{cases} 0 \text{ if } m = j \\ X_m(t) \;\; \text{otherwise.} \end{cases}$$

Using the fundamental theorem of calculus we obtain that

$$
\begin{aligned}
&\Big|\mathbb{E}(f_\epsilon(\sum_j X_j) - \mathbb{E}(f_\epsilon(\sum_j Z_j))\Big| \\
&\leq \int_0^1 \Big|\partial_t \mathbb{E}(f_\epsilon(\sum_j X_j(t)))\Big| dt \\
&\overset{(d_1)}{\leq} \int_0^1 \Big|\mathbb{E}\Big(f_\epsilon'(\sum_j X_j(t)) \sum_j \frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}\Big)\Big| dt \\
&\overset{(d_2)}{\leq} \int_0^1 \sum_{j\leq n} \Big|\mathbb{E}\Big(f_\epsilon'(\sum X^{j,0}(t))[\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}]\Big)\Big| dt \\
&+ \int_0^1 \sum_{j\leq n} \Big|\mathbb{E}\Big(f_\epsilon''(\sum X^{j,0}(t))[\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}][X_j\sqrt{t} + Z_j\sqrt{1-t}]\Big)\Big| dt \\
&+ \int_0^1 \sum_{j\leq n} \Big|\mathbb{E}\Big(f_\epsilon^{(3)}(\sum_j \tilde{X}_j(t))[\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}][X_j\sqrt{t} + Z_j\sqrt{1-t}]^2\Big)\Big| dt \\
&\leq (a) + (b) + (c)
\end{aligned}
$$

where $\sum \tilde{X}_j(t) \in [\sum X_j(t), \sum X^{j,0}(t)]$ and where d_1 is a result of the product law and d_2 of Taylor's expansion. Using the independence between $X^{j,0}(t)$ and X_j and Z_j and the fact that those latter are centered, we obtain that $(a) = 0$. Similarly, we notice for all $j \leq n$ that

$$
\begin{aligned}
&\mathbb{E}\Big(f_\epsilon''(\sum X^{j,0}(t))[\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}][X_j\sqrt{t} + Z_j\sqrt{1-t}]\Big) \\
&= \mathbb{E}\Big(f_\epsilon''(\sum X^{j,0}(t))\Big)\mathbb{E}\Big([\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}][X_j\sqrt{t} + Z_j\sqrt{1-t}]\Big) \\
&\leq \mathbb{E}\Big(f_\epsilon''(\sum X^{j,0}(t))\Big)\mathbb{E}\Big([\frac{X_j^2}{2} - \frac{Z_j^2}{2}]\Big) = 0.
\end{aligned}
$$

Hence $(b) = 0$.

Finally we can note that $\|f_\epsilon^{(3)}\| \leq \frac{4}{\epsilon^2}$. Hence, thanks to Jensen inequality we know that there exists absolute constants $C, C_2 > 0$ such that

$$
\begin{aligned}
&\Big|\mathbb{E}\Big(f_\epsilon^{(3)}(\sum_j \tilde{X}_j(t))[\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}][X_j\sqrt{t} + Z_j\sqrt{1-t}]^2\Big)\Big| \\
&\leq \frac{4}{\epsilon^2}\mathbb{E}\Big(\Big|\frac{X_j}{2\sqrt{t}} - \frac{Z_j}{2\sqrt{1-t}}\Big|[X_j\sqrt{t} + Z_j\sqrt{1-t}]^2\Big)\Big| \\
&\leq \frac{C}{\epsilon^2}\max(\|X_j\|_3^3, \|Z_j\|_3^3)\Big[\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{1-t}}\Big] \\
&\leq \frac{C}{\epsilon^2}\max(\|X_j\|_3^3,\ \sqrt{3}^3\|X_j\|_2^3)\Big[\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{1-t}}\Big] \\
&\leq \frac{C_2}{\epsilon^2}\|X_j\|_3^3\Big[\frac{1}{\sqrt{t}} + \frac{1}{\sqrt{1-t}}\Big]
\end{aligned}
$$

where we used the fact that if $Z \sim \mathcal{N}(0,1)$ then $\|Z\|_3 \leq \sqrt{3}$ coupled with the fact that $\|Z_j\|_2 = \|X_j\|_2$. Hence we obtain that there is an absolute constant $\kappa > 0$ such that

$$(c) \leq \frac{\kappa n}{\epsilon^2} \max_{j \leq n} \|X_j\|_3^3 \tag{95}$$

This gives us the desired result by choosing $\epsilon := \left(\kappa n\right)^{1/3} \max_{j \leq n} \|X_j\|_3$. ■

K.5. Asymptotic normality under local dependency assumption

Lemma 42 *Suppose that (X_i) are centered random vectors such that the array $(X_{i,\ell})_{i,\ell}$ is locally dependent. Write $B_{i,\ell}$ the dependency neighborhood of the entry $X_{i,\ell}$. For a fixed $q \in \mathbb{N}$, let $\mathcal{F}_q$ be the class of $\mathbb{R}^q \to \mathbb{R}$ continuously differentiable functions with $\|f\|_\infty \leq 1$ and $\| \|\partial f\| \|_\infty \leq 1$. Then there is a constant $C_q > 0$ that depends only on q such that*

$$\sup_{\substack{f \in \mathcal{F}_q \\ \theta_1,\ldots,\theta_q \in \mathcal{S}^{k-1}}} \sup_{\substack{(\beta_{11},\ldots,\beta_{kq}) \\ \in \mathcal{S}_p^{kq}}} \left| \mathbb{E} f \begin{pmatrix} \sum_{i=1}^k \theta_{1i} X_i^\intercal \beta_{1i} \\ \vdots \\ \sum_{i=1}^k \theta_{qi} X_i^\intercal \beta_{qi} \end{pmatrix} - \mathbb{E} f \begin{pmatrix} \sum_{i=1}^k \theta_{1i} G_i^\intercal \beta_{1i} \\ \vdots \\ \sum_{i=1}^k \theta_{qi} G_i^\intercal \beta_{qi} \end{pmatrix} \right|$$
$$\leq C_q \left(k^2 \, p^{3/2-r} \max_{i,\ell} |B_{i,\ell}|^{3/2} \max_{i,\ell} \|X_{i,\ell}\|_{L_3}^3 \right)^{1/(2q+1)} , \tag{96}$$

where $\mathcal{S}^{k-1}$ denotes the unit sphere in $\mathbb{R}^k$ and (G_i) is an independent sequence of mean-zero Gaussian vectors chosen such that cov$(G_{i,l}, G_{j,m})$ = cov$(X_{i,l}, X_{j,m})$.

Proof Fix $f \in \mathcal{F}_q$. Let $\epsilon > 0$ and define a smoothed version of f,

$$f_\epsilon(u) := \frac{1}{(2\epsilon)^{2q}} \int_{u_1-\epsilon}^{u_1+\epsilon} \cdots \int_{u_q-\epsilon}^{u_q+\epsilon} \int_{t_1-\epsilon}^{t_1+\epsilon} \cdots \int_{t_q-\epsilon}^{t_q+\epsilon} f(y) \, dy_1 \ldots dy_q \, dt_1 \ldots dt_q \ .$$

Note that f_ϵ is thrice differentiable and, as f is Lipschitz, we have $\sup_x |f_\epsilon(x) - f(x)| \leq 6\epsilon \sqrt{q}$. Write

$$X_j(t) := \sqrt{t} X_j + \sqrt{1-t} G_j \ , \ \ a_{jl} := (\theta_{1j}\beta_{1jl} \, , \, \ldots \, , \, \theta_{qj}\beta_{qjl}) \in \mathbb{R}^q \ , \ \ A_j := \begin{pmatrix} \leftarrow & a_{j1}^\intercal & \rightarrow \\ & \vdots & \\ \leftarrow & a_{jp}^\intercal & \rightarrow \end{pmatrix} \in \mathbb{R}^{p \times q} \ ,$$

and

$$(X_{j,i,\ell,0}(t))_m := \begin{cases} 0 \text{ if } (j,m) \in B_{i,\ell} \\ X_{j,m}(t) \ \text{ otherwise.} \end{cases} \ .$$

Using the fundamental theorem of calculus we obtain that

$$\begin{aligned}
&\left|\mathbb{E}f_\epsilon\begin{pmatrix}\sum_{i=1}^k \theta_{1i}X_i^\intercal\beta_{1i}\\ \vdots \\ \sum_{i=1}^k \theta_{qi}X_i^\intercal\beta_{qi}\end{pmatrix} - \mathbb{E}f_\epsilon\begin{pmatrix}\sum_{i=1}^k \theta_{1i}G_i^\intercal\beta_{1i}\\ \vdots \\ \sum_{i=1}^k \theta_{qi}G_i^\intercal\beta_{qi}\end{pmatrix}\right| \;=\; \Big|\mathbb{E}\Big[f_\epsilon\Big(\sum_{j\le n} X_j^\intercal A_j\Big)\Big] - \mathbb{E}\Big[f_\epsilon\Big(\sum_{j\le n} G_j^\intercal A_j\Big)\Big]\Big| \\
&\le \int_0^1 \Big|\partial_t\,\mathbb{E}\Big[f_\epsilon\Big(\sum_j\; X_j^\intercal(t)\,A_j\Big)\Big]\Big|\,dt \\
&\overset{(d_1)}{\le} \int_0^1 \Big|\mathbb{E}\Big[\partial f_\epsilon\big(\sum_j X_j^\intercal(t)A_j\big)\sum_{i\le k} A_i^\intercal\Big(\frac{X_i}{2\sqrt{t}} - \frac{G_i}{2\sqrt{1-t}}\Big)\Big]\Big|dt \\
&\overset{(d_2)}{\le} \int_0^1 \Big|\sum_{i\le k}\sum_{\ell\le p}\mathbb{E}\Big[\partial f_\epsilon\Big(\sum_j X_{j,i,\ell,0}^\intercal(t)A_j\Big)^\intercal a_{il}\Big(\frac{X_{i,\ell}}{2\sqrt{t}} - \frac{G_{i,\ell}}{2\sqrt{1-t}}\Big)\Big]\Big|dt \\
&+ \frac{1}{2}\int_0^1 \Big|\sum_{i\le k}\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell})\in B_{i,\ell}}\mathbb{E}\Big[(X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t})a_{\tilde{i},\tilde{l}}^\intercal\,\partial^2 f_\epsilon\Big(\sum_j X_{j,i,\ell,0}^T(t)A_j\Big)\,a_{il} \\
&\qquad\qquad\qquad\qquad\qquad\qquad\Big(\frac{X_{i,l}}{2\sqrt{t}} - \frac{G_{i,l}}{2\sqrt{1-t}}\Big)\Big]\Big|dt \\
&+ \frac{1}{6\epsilon^q}\int_0^1 \mathbb{E}\Big[\Big\|\sum_{i\le k}\sum_{\ell\le p}\sum_{\substack{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\\ \in B_{i,\ell}}}\Big(\frac{X_{i,l}}{2\sqrt{t}} - \frac{G_{i,l}}{2\sqrt{1-t}}\Big)(X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t}) \\
&\qquad\qquad\qquad\qquad\qquad\qquad(X_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{t} + G_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{1-t})\,(a_{il}\otimes a_{\tilde{i}\tilde{\ell}}\otimes a_{\tilde{i}_2,\tilde{\ell}_2})\Big\|\,\Big]\,dt \\
&\le (a) + (b) + (c)\;,
\end{aligned}$$

where $\tilde{X}_{j,i,l}(t) \in [X_j(t), X_{j,i,l,0}(t)]$. In (d_1), we have used the product rule; in (d_2), we have used a third-order Taylor expansion together with the bound that $\|\partial^3 f_\epsilon\|_\infty \le \frac{1}{\epsilon^q}$. Using the independence between $X_{j,i,\ell,0}(t)$ and $X_{i,\ell}$ and $G_{i,\ell}$ and the fact that these variables are centered, we obtain that for all $i \le k$ and $\ell \le p$ we have

$$\mathbb{E}\Big[\partial f_\epsilon\Big(\sum_j X_{j,i,\ell,0}^\intercal(t)\Theta_j^\beta\Big)(\Theta_i^\beta)^\intercal\Big(\frac{X_{i,\ell}}{2\sqrt{t}} - \frac{G_{i,\ell}}{2\sqrt{1-t}}\Big)\Big] \;=\; 0\;.$$

Hence we know that $(a) = 0$. Similarly we notice that

$$\mathbb{E}\Big[(X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t})(\Theta_{\tilde{i}}^{\beta,(\tilde{l})})^\intercal\,\partial^2 f_\epsilon\Big(\sum_j X_{j,i,\ell,0}^T(t)\Theta_j^\beta\Big)\,(\Theta_i^\beta)^\intercal\Big(\frac{X_i}{2\sqrt{t}} - \frac{G_i}{2\sqrt{1-t}}\Big)\Big] \;=\; 0\;,$$

where we use the independence between $(X_{i,j})$ and $(G_{i,j})$ to ignore cross terms and the fact that $\mathbb{E}(X_{i,\ell}X_{\tilde{i},\tilde{\ell}}) = \mathbb{E}(G_{i,\ell}G_{\tilde{i},\tilde{\ell}})$. Hence $(b) = 0$. Finally to handle (c), we see that

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\sum_{i\le k}\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}\Big(\frac{X_{i,l}}{2\sqrt{t}} - \frac{G_{i,l}}{2\sqrt{1-t}}\Big) \\
&\qquad\qquad \times (X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t})\,(X_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{t} + G_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{1-t})\,(a_{il}\otimes a_{\tilde{i}\tilde{\ell}}\otimes a_{\tilde{i}_2,\tilde{\ell}_2})\Big\|\Big] \\
&= \mathbb{E}\Big[\Big(\textstyle\sum_{s,\tilde{s},\tilde{s}_2\le q}\Big(\displaystyle\sum_{i\le k}\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}\Big(\frac{\theta_{si}\beta_{sil}X_{i,l}}{2\sqrt{t}} - \frac{\theta_{si}\beta_{sil}G_{i,l}}{2\sqrt{1-t}}\Big)\theta_{\tilde{s}\tilde{i}}\beta_{\tilde{s}\tilde{i}\tilde{l}}(X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t}) \\
&\qquad\qquad \theta_{\tilde{s}_2\tilde{i}_2}\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}(X_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{t} + G_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{1-t})\Big)^2\Big)^{1/2}\Big] \\
&\overset{(d_3)}{\le} \textstyle\sum_{s,\tilde{s},\tilde{s}_2\le q}\sum_{i\le k}\mathbb{E}\Big[\displaystyle\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\theta_{si}\beta_{sil}|\Big|\frac{X_{i,l}}{2\sqrt{t}} - \frac{G_{i,l}}{2\sqrt{1-t}}\Big|\,|\theta_{\tilde{s}\tilde{i}}\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,\big|X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t}\big| \\
&\qquad\qquad |\theta_{\tilde{s}_2\tilde{i}_2}\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}|\,\big|X_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{t} + G_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{1-t}\big|\Big] \\
&\overset{(d_4)}{\le} \textstyle\sum_{s,\tilde{s},\tilde{s}_2\le q}\sum_{i\le k}\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\beta_{sil}|\,|\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,|\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}| \\
&\qquad \times\mathbb{E}\Big[\Big|\frac{X_{i,l}}{2\sqrt{t}} - \frac{G_{i,l}}{2\sqrt{1-t}}\Big|\,\big|X_{\tilde{i},\tilde{\ell}}\sqrt{t} + G_{\tilde{i},\tilde{\ell}}\sqrt{1-t}\big|\big|X_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{t} + G_{\tilde{i}_2,\tilde{\ell}_2}\sqrt{1-t}\big|\Big] \\
&\overset{(d_5)}{\le} \textstyle\sum_{s,\tilde{s},\tilde{s}_2\le q}\sum_{i\le k}\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\beta_{sil}|\,|\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,|\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}| \times\Big(\displaystyle\frac{1}{2\sqrt{t}} + \frac{1}{2\sqrt{1-t}}\Big)(1+\sqrt{3})^2\max_{i,\ell}\|X_{i,\ell}\|_{L_3}^3\,.
\end{aligned}$$

In (d_3), we have moved the summations outside a squareroot and an absolute value; in (d_4), we have noted that $\theta \in \mathcal{S}^{k-1}$; in (d_5), we have used that for $Z \sim \mathcal{N}(0,1)$, $\|Z\|_{L_3} \le \sqrt{3}$. Now let $(\beta_{\mathrm{mix}})_s := (|(\beta_{\mathrm{mix}})_{s11}|,\ldots,|(\beta_{\mathrm{mix}})_{skp}|) \in \mathbb{R}^{kp}$ and $M^{(i,\ell)} \in \mathbb{R}^{kp\times kp}$ be a matrix with entries $M^{(i,\ell)}_{(i',\ell'),(i'',\ell'')} = \mathbb{I}\{(i',\ell')\in B_{i,\ell}\}\,\mathbb{I}\{(i'',\ell'')\in B_{i,\ell}\}$. Also recall that by the definition of $\mathcal{S}_p$ in (4), there are fixed constants $L, r > 0$ such that $\|\beta_{si}\|_\infty \le Lp^{1/2-r}$ and $\|\beta_{si}\|_2 \le Lp^{1/2}$ for all $s \in q, i \le k$. Then

$$\begin{aligned}
\textstyle\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\beta_{sil}|\,|\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,|\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}| &\le Lp^{1/2-r}\textstyle\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,|\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}| \\
&= Lp^{1/2-r}\textstyle\sum_{\ell\le p}(\beta_{\mathrm{mix}})_{\tilde{s}}^\intercal M^{(i,\ell)}(\beta_{\mathrm{mix}})_{\tilde{s}_2} \\
&\le Lp^{1/2-r}\,\|(\beta_{\mathrm{mix}})_{\tilde{s}}\|\,\|(\beta_{\mathrm{mix}})_{\tilde{s}_2}\|\,\Big\|\textstyle\sum_{\ell\le p}M^{(i,\ell)}\Big\|_{op} \\
&\le L^3kp^{3/2-r}\Big\|\textstyle\sum_{\ell\le p}M^{(i,\ell)}\Big\|_{op}
\end{aligned}$$

In the last line, we have noted that $\|(\beta_{\mathrm{mix}})_{\tilde{s}}\|^2 = \sum_{\tilde{i}\le k}\|(\beta_{\mathrm{mix}})_{\tilde{s}\tilde{i}}\|^2 \le L^2kp$. Now observe that the (i',ℓ')-th column of the matrix $\sum_{\ell\le p}M^{(i,\ell)}$ is given by

$$\Big(\textstyle\sum_{\ell\le p}\mathbb{I}\{(i',\ell')\in B_{i,\ell}\}\,\mathbb{I}\{(i'',\ell'')\in B_{i,\ell}\}\Big)_{i''\le k,\ell''\le p}\,.$$

Since $|B_{i,\ell}| \le \max_{i,\ell}|B_{i,\ell}|$, the column has at most $\max_{i,\ell}|B_{i,\ell}|$ non-zero entries. For each (i'',ℓ''), since the dependency neighborhood induces an equivalence relation and $|B_{i'',\ell''}| \le \max_{i,\ell}|B_{i,\ell}|$, the (i'',ℓ'')-th entry cannot exceed $\max_{i,\ell}|B_{i,\ell}|$. In other words, the l_2-norm of each column vector of $\sum_{\ell\le p}M^{(i,\ell)}$ cannot exceed $\sqrt{\max_{i,\ell}|B_{i,\ell}|\times\max_{i,\ell}|B_{i,\ell}|^2} = \max_{i,\ell}|B_{i,\ell}|^{3/2}$, which implies

$$\textstyle\sum_{\ell\le p}\sum_{(\tilde{i},\tilde{\ell}),(\tilde{i}_2,\tilde{\ell}_2)\in B_{i,\ell}}|\beta_{sil}|\,|\beta_{\tilde{s}\tilde{i}\tilde{l}}|\,|\beta_{\tilde{s}_2\tilde{i}_2\tilde{l}_2}| \le L^3kp^{3/2-r}\displaystyle\max_{i,\ell}|B_{i,\ell}|^{3/2}\;;.$$

Combining the bounds, we obtain that for some constant $C_q > 0$ depending only on q,

$$(c) \leq \frac{C_q}{\epsilon^{2q}} k^2 \, p^{3/2-r} \max_{i,\ell} |B_{i,\ell}|^{3/2} \max_{i,\ell} \|X_{i,\ell}\|_{L_3}^3 .$$

This gives us the desired result by choosing $\epsilon := (C_q \, k^2 \, p^{3/2-r} \, \max_{i,\ell} |B_{i,\ell}|^{3/2} \, \max_{i,\ell} \|X_{i,\ell}\|_{L_3}^3)^{1/(2q+1)}$.
■

K.6. Polynomial Approximation Properties

In this section we discuss some of the properties of our polynomial approximation that are used in the proof of our main theorem and also Theorem 27.

Lemma 43 *Let $\alpha, \delta, \gamma, \tau > 0$. Then there exists finite $\mathsf{D} = \mathsf{D}(k, \alpha, \tau)$ such that, if we define*

$$Q_{\mathsf{D}}(x) := \sum_{\ell=0}^{\mathsf{D}} (1-x)^{\ell}, \qquad R_{\mathsf{D}}(x) := \frac{1}{x} - Q_{\mathsf{D}}(x),$$

then

$$\mathbb{E}_{(i,k)} \left[R_{\mathsf{D}} \left(\langle e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \rangle_{i,k} \right)^2 \right] < \tau.$$

Proof For $t > 0$, define the event

$$\mathcal{A}_t := \left\{ \max_{j \in \mathcal{B}_i} \left| U_j^{\mathsf{T}} \beta \right| \leq t \right\} = \bigcap_{j \in \mathcal{B}_i} \left\{ \left| U_j^{\mathsf{T}} \beta \right| \leq t \right\}.$$

Then we have that

$$\begin{aligned} \mathbb{P}(\mathcal{A}_t^c) &\overset{(i)}{=} \mathbb{P} \left(\bigcup_{j \in \mathcal{B}_i} \left\{ \left| U_j^{\mathsf{T}} \beta \right| > t \right\} \right) \\ &\overset{(ii)}{\leq} \sum_{j \in \mathcal{B}_i} \mathbb{P} \left(\left| U_j^{\mathsf{T}} \beta \right| > t \right) \\ &\overset{(iii)}{\leq} \sum_{j \in \mathcal{B}_i} 2 e^{-ct^2/\mathsf{C}_1^2} \\ &\leq \mathsf{C}_2 k e^{-ct^2}. \end{aligned} \tag{97}$$

where (i) is via De Morgan's Law, (ii) is via a union bound, and (iii) is from (60) which bounds the sub-Gaussian norm of each $\left| U_j^{\mathsf{T}} \beta \right|$ along with Proposition 2.5.2 of Vershynin (2018). We can then say that

$$\begin{aligned} \mathbb{E}_{(i,k)} \left[R_{\mathsf{D}} \left(\langle e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \rangle_{i,k} \right)^2 \right] &\overset{(i)}{\leq} \mathbb{E}_{(i,k)} \left\langle R_{\mathsf{D}} \left(e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \right)^2 \right\rangle_{i,k} \\ &\overset{(ii)}{=} \left\langle \mathbb{E}_{(i,k)} \left[R_{\mathsf{D}} \left(e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \right)^2 \right] \right\rangle_{i,k} \\ &\overset{(iii)}{=} \left\langle \mathbb{E}_{(i,k)} \left[R_{\mathsf{D}} \left(e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \right)^2 \mathbb{I}_{\mathcal{A}_t} \right] \right\rangle_{i,k} \qquad (98) \\ &\quad + \left\langle \mathbb{E}_{(i,k)} \left[R_{\mathsf{D}} \left(e^{-\alpha \sum_{j \in \mathcal{B}_i} \omega_j \ell(\eta_j, U_j^{\mathsf{T}} \beta)} \right)^2 \mathbb{I}_{\mathcal{A}_t^c} \right] \right\rangle_{i,k} \qquad (99) \end{aligned}$$

where (i) is because R_{D}^2 is convex, as it is the square of a positive, convex function, (ii) is because $\mathbb{E}_{(i,k)}[\ ]$ and $\langle\ \rangle_{i,k}$ commute, and (iii) is the Law of Total Probability. We first bound the term inside the expectation of (99) as

$$\begin{aligned}\mathbb{E}_{(i,k)}\left[R_{\mathsf{D}}\left(e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\right)^2\mathbb{I}_{\mathcal{A}_t^c}\right] &\overset{(i)}{\leq} \mathbb{E}\left[R_{\mathsf{D}}\left(e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\right)^4\right]^{1/2}\mathbb{P}\left(\mathcal{A}_t^c\right)^{1/2} \\ &\overset{(ii)}{\leq} \mathbb{E}\left[e^{4\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\right]^{1/2}\mathsf{C}_3\sqrt{k}e^{-ct^2} \\ &\overset{(iii)}{\leq} e^{\mathsf{C}_4k^2\alpha^2}\mathsf{C}_3\sqrt{k}e^{-ct^2} \\ &= \mathsf{C}(k,\alpha)e^{-ct^2},\end{aligned} \tag{100}$$

where (i) is via Cauchy-Schwarz and block dependence, (ii) is via (97) and the fact that

$$Q_{\mathsf{D}}(x) > 0 \implies R_{\mathsf{D}}(x) = \frac{1}{x} - Q_{\mathsf{D}}(x) < \frac{1}{x}$$

for $x \in (0,1)$, and (iii) is via (62). Thus, if we choose t sufficiently large, namely

$$t > \sqrt{\frac{1}{c}\log\left(\frac{2\mathsf{C}(k,\alpha)}{\tau}\right)},$$

then (100) yields that

$$\mathbb{E}_{(i,k)}\left[R_{\mathsf{D}}\left(e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\right)^2\mathbb{I}_{\mathcal{A}_t^c}\right] \leq \frac{\tau}{2}. \tag{101}$$

For (98), we know that since the event $\mathcal{A}_t$ occurs in this case, we have

$$\mathcal{A}_t \implies \sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta) \leq \sum_{j\in\mathcal{B}_i}\left|U_j^\intercal\beta\right| + 1 \leq k(t+1),$$

and so this forces that the argument of R_{D} satisfies

$$\exp\left(-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell_j(\beta)\right) \in [e^{-\alpha k(t+1)}, 1].$$

By definition of $Q_{\mathsf{D}}(x)$ being the power series of $\frac{1}{x}$ with radius of convergence equal to 1, there must exist $\mathsf{D}(k,\alpha,\tau)$ such that

$$\sup_{x\in[e^{-\alpha k(t+1)},1]}|R_{\mathsf{D}}(x)| \leq \sqrt{\frac{\tau}{2}}.$$

This means that the term inside the expectation of (98) may be bounded as

$$\mathbb{E}_{(i,k)}\left[R_{\mathsf{D}}\left(e^{-\alpha\sum_{j\in\mathcal{B}_i}\omega_j\ell(\eta_j,U_j^\intercal\beta)}\right)^2\mathbb{I}_{\mathcal{A}_t}\right] \leq \frac{\tau}{2}, \tag{102}$$

and so the result follows from this choice of D by combining (101) and (102).

■

K.7. Properties of sub-Gaussian Vectors

Lemma 44 *Let Y be a sub-Gaussian vector in $\mathbb{R}^d$ with sub-Gaussian constant K. Write $\Sigma_Y :=$ Var(Y). Then there exists $\mathsf{C} > 0$ such that*

$$\|\Sigma_Y\|_{\mathrm{op}} \leq \mathsf{C}\mathsf{K}^2 .$$

Proof Define $Z := Y - \mathbb{E}[Y]$, which by Lemma 2.6.8. of Vershynin (2018) is still sub-Gaussian with

$$\|Z\|_{\psi_2} \leq \mathsf{C}_1\mathsf{K}$$

for some fixed $\mathsf{C}_1 > 0$. Now, let $v \in \mathbb{R}^d$. We first know by Definition 9 that since Z is sub-Gaussian with constant $\mathsf{C}_1\mathsf{K}$, $Z^\top v$ must also be sub-Gaussian with constant at most $\mathsf{C}_1\mathsf{K}\|v\|$. We observe that

$$v^\top \Sigma_Y v \overset{(i)}{=} \mathrm{Var}(Z^\top v) \overset{(ii)}{\leq} \mathsf{C}_2\mathsf{K}^2\|v\|^2 ,$$

where (i) is because Y and Z share the same covariance matrix, and the inequality in (ii) is via Proposition 2.5.2 of Vershynin (2018). This lets us conclude that

$$\frac{v^\top \Sigma_Y v}{\|v\|^2} \leq \mathsf{C}_2\mathsf{K}^2 ,$$

and since this holds for all $v \in \mathbb{R}^d$, it holds for the supremum, which exactly defines the operator norm as Σ_Y is necessarily positive semi-definite. ■

Appendix L. Proofs for the dependent CGMT

In this section, we prove Theorem 13, which recovers Theorem 5 directly, and Corollary 7. The proof recipe is similar to that of a standard CGMT: We start by proving a Gaussian min-max theorem (GMT) on discrete sets in Lemma 45, proceed to extend it to compact sets in Lemma 46, and then prove the results in Theorem 13. Corollary 7 then follows directly from Theorem 13(ii).

As with the standard CGMT, the Gaussian min-max theorem (GMT) on discrete sets is proved for a surrogate optimization problem. Let $(\xi_l)_{l\leq M}$ be a collection of univariate standard Gaussians independent of $\mathbf{H}$, and define

$$\Psi^{\xi}_{\mathcal{S}_w,\mathcal{S}_u} := \min_{w\in\mathcal{S}_w} \max_{u\in\mathcal{S}_u} L^{\xi}_{\Psi}(w,u) , \qquad \text{where} \quad L^{\xi}_{\Psi}(w,u) := w^\top \mathbf{H} u + \textstyle\sum_{l=1}^{M} \xi_l \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} + f(w,u) .$$

We also recall the risk $\psi_{\mathcal{I}_p,\mathcal{I}_n}$ of the auxiliary optimization defined in Theorem 13.

Lemma 45 (GMT on discrete sets) *Let $\mathcal{I}_p \subseteq \mathbb{R}^p$, $\mathcal{I}_n \subseteq \mathbb{R}^n$ be discrete sets, and f be finite on $\mathcal{I}_p \times \mathcal{I}_n$. Then for all $c \in \mathbb{R}$,*

$$\mathbb{P}(\Psi^{\xi}_{\mathcal{I}_p,\mathcal{I}_n} \geq c) \;\geq\; \mathbb{P}(\psi_{\mathcal{I}_p,\mathcal{I}_n} \geq c) .$$

Proof of Lemma 45 Similar to the proof for the standard GMT (see e.g. proof of Lemma A.1.1 of Thrampoulidis (2016)), the proof relies on an application of Gordon's Gaussian comparison inequality (see e.g. Corollary 3.13 of Ledoux and Talagrand (1991)) applied to two suitably defined Gaussian processes. Consider the two centred Gaussian processes indexed on the set $\mathcal{I}_p \times \mathcal{I}_n$:

$$\begin{aligned} Y_{w,u} &:= w^\top \mathbf{H} u + \textstyle\sum_{l=1}^M \xi_l \, \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \;, \\ X_{w,u} &:= \textstyle\sum_{l=1}^M \big(\|w\|_{\Sigma^{(l)}} \mathbf{h}_l^\top (\tilde{\Sigma}^{(l)})^{1/2} u + w^\top (\Sigma^{(l)})^{1/2} \mathbf{g}_l \|u\|_{\tilde{\Sigma}^{(l)}} \big) \;. \end{aligned}$$

To compare their second moments, we use the independence of $\mathbf{H}$ and $\{\xi_l\}_{l \leq M}$ as well as the independence of $(\mathbf{h}_l, \mathbf{g}_l)_{l \leq M}$: For $w, w' \in \mathcal{I}_p$ and $u, u' \in \mathcal{I}_n$, we have

$$\begin{aligned} \mathbb{E}[Y_{w,u} Y_{w',u'}] - \mathbb{E}[X_{w,u} X_{w',u'}] &\overset{(a)}{=} \mathbb{E}[w^\top \mathbf{H} u (w')^\top \mathbf{H} u'] + \textstyle\sum_{l=1}^M \|w\|_{\Sigma^{(l)}} \, \|w'\|_{\Sigma^{(l)}} \, \|u\|_{\tilde{\Sigma}^{(l)}} \, \|u'\|_{\tilde{\Sigma}^{(l)}} \\ &\quad - \textstyle\sum_{l=1}^M \big(\|w\|_{\Sigma^{(l)}} \, \|w'\|_{\Sigma^{(l)}} \, u^\top \tilde{\Sigma}^{(l)} u' + w^\top \Sigma^{(l)} w' \, \|u\|_{\tilde{\Sigma}^{(l)}} \, \|u'\|_{\tilde{\Sigma}^{(l)}} \big) \\ &\overset{(b)}{=} \textstyle\sum_{l=1}^M \Big(w^\top \Sigma^{(l)} w' \, u^\top \tilde{\Sigma}^{(l)} u' + \|w\|_{\Sigma^{(l)}} \, \|w'\|_{\Sigma^{(l)}} \, \|u\|_{\tilde{\Sigma}^{(l)}} \, \|u'\|_{\tilde{\Sigma}^{(l)}} \\ &\qquad\qquad - \|w\|_{\Sigma^{(l)}} \, \|w'\|_{\Sigma^{(l)}} \, u^\top \tilde{\Sigma}^{(l)} u' \; - \; w^\top \Sigma^{(l)} w' \|u\|_{\tilde{\Sigma}^{(l)}} \|u'\|_{\tilde{\Sigma}^{(l)}} \Big) \\ &= \textstyle\sum_{l=1}^M \big(\|w\|_{\Sigma^{(l)}} \, \|w'\|_{\Sigma^{(l)}} - w^\top \Sigma^{(l)} w' \big) \big(\|u\|_{\tilde{\Sigma}^{(l)}} \, \|u'\|_{\tilde{\Sigma}^{(l)}} - u^\top \tilde{\Sigma}^{(l)} u' \big) \;. \end{aligned} \tag{103}$$

In (a), we have used that ξ_l's, $\mathbf{h}_l$'s and $\mathbf{g}_l$'s are all standard Gaussians; in (b), we have used

$$\begin{aligned} \mathbb{E}[w^\top \mathbf{H} u (w')^\top \mathbf{H} u'] &= \textstyle\sum_{i,i'=1}^n \sum_{j,j'=1}^p \; w_i w'_{i'} u_j u'_{j'} \, \mathbb{E}[H_{ij} H_{i'j'}] \\ &= \textstyle\sum_{l=1}^M \sum_{i,i'=1}^n \sum_{j,j'=1}^p \; w_i \Sigma^{(l)}_{ii'} w'_{i'} u_j \tilde{\Sigma}^{(l)}_{jj'} u'_{j'} \\ &= \textstyle\sum_{l=1}^M w^\top \Sigma^{(l)} w' \, u^\top \tilde{\Sigma}^{(l)} u' \;. \end{aligned}$$

By the positive semi-definiteness of $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$, (103) is non-negative, and equals to zero when $w = w'$. This shows that the Gaussian processes $(Y_{w,u})_{w \in \mathcal{I}_p, u \in \mathcal{I}_n}$ and $(X_{w,u})_{w \in \mathcal{I}_p, u \in \mathcal{I}_n}$ verify the conditions of the Gaussian comparison inequality (Corollary 3.13 of Ledoux and Talagrand (1991)) and therefore for any real sequence $(\lambda_{w,u})_{w \in \mathcal{I}_p, u \in \mathcal{I}_n}$,

$$\mathbb{P}\big(\cap_{w \in \mathcal{I}_p} \cup_{v \in \mathcal{I}_n} \{ Y_{w,u} \geq \lambda_{w,u} \} \big) \;\geq\; \mathbb{P}\big(\cap_{w \in \mathcal{I}_p} \cup_{v \in \mathcal{I}_n} \{ X_{w,u} \geq \lambda_{w,u} \} \big) \;.$$

Choosing $\lambda_{w,u} = -f(w,u) + c$ yields that

$$\mathbb{P}\Big(\min_{w \in \mathcal{I}_p} \max_{v \in \mathcal{I}_n} (Y_{w,u} + f(w,u)) \geq c \Big) \;\geq\; \mathbb{P}\Big(\min_{w \in \mathcal{I}_p} \max_{v \in \mathcal{I}_n} (X_{w,u} + f(w,u)) \geq c \Big) \;.$$

Noting that the two min-max quantities correspond to $\Psi^{\xi}_{\mathcal{I}_p, \mathcal{I}_n}$ and $\psi_{\mathcal{I}_p, \mathcal{I}_n}$ concludes the proof. ■

The next result extends Lemma 45 to compact sets.

Lemma 46 (GMT for compact sets) *Suppose $\mathcal{S}_w \subset \mathbb{R}^p$ and $\mathcal{S}_u \subset \mathbb{R}^n$ are compact and f is continuous on $\mathcal{S}_w \times \mathcal{S}_u$. Then for all $c \in \mathbb{R}$,*

$$\mathbb{P}(\Psi^{\xi}_{\mathcal{S}_w, \mathcal{S}_u} \geq c) \;\geq\; \mathbb{P}(\psi_{\mathcal{S}_p, \mathcal{S}_n} \geq c) \;.$$

Proof of Lemma 46 The proof is almost identical to the proof of standard GMT results for compact sets, now that we have established Lemma 45: We show by a compactness argument that both losses only change a little when replacing $\mathcal{S}_w$ and $\mathcal{S}_u$ by their δ-nets $\mathcal{S}_p^\delta$ and $\mathcal{S}_n^\delta$, induced by the Euclidean norms on $\mathbb{R}^n$ and $\mathbb{R}^d$ respectively. The only difference from their proof is that we use a slightly different concentration inequality. Therefore we only set up the essential notation, highlight the differences and refer interested readers to the proof of Theorem 3.2.1 of Thrampoulidis (2016), found in Pg 185-187.

First fix some $\epsilon > 0$. Since f is continuous and thereby uniformly continuous on the compact set $\mathcal{S}_p^\delta \times \mathcal{S}_n^\delta$, there exists some $\delta = \delta(\epsilon) > 0$ such that for all $(w,u),(w',u') \in \mathcal{S}_w \times \mathcal{S}_u$ with $\|(w,u)-(w',u')\| \leq \delta$, we have $\|f(w,u)-f(w',u')\| \leq \epsilon$. Use this δ to form the δ-nets $\mathcal{S}_p^\delta$ and $\mathcal{S}_n^\delta$. We also write $\|\bullet\|_{op}$ as the operator norm of a matrix, and write

$$S \;:=\; \max\nolimits_{1\leq l\leq M} \max\{\|\Sigma^{(l)}\|_{\text{op}}\,,\, \|\tilde{\Sigma}^{(l)}\|_{\text{op}}\} \qquad \text{and} \qquad K \;:=\; \max\{\, \sup\nolimits_{w\in\mathcal{S}_w} \|w\|\,,\, \sup\nolimits_{u\in\mathcal{S}_u} \|u\|\}\;.$$

K is bounded since $\mathcal{S}_w$ and $\mathcal{S}_u$ are compact, and for $w \in \mathcal{S}_w$, $u \in \mathcal{S}_u$ and $l \leq M$, we have

$$\|w\| \;\leq\; K\,, \qquad \|w\|_{\Sigma^{(l)}} \;\leq\; S\,K\,, \qquad \|u\| \;\leq\; K\,, \qquad \|u\|_{\tilde{\Sigma}^{(l)}} \;\leq\; S\,K\,.$$

Then by the same argument as the proof of Theorem 3.2.1 of Thrampoulidis (2016), there exists $w_1 \in \mathcal{S}_w$, $w_1' \in \mathcal{S}_p^\delta$ with $\|w_1 - w_1'\| \leq \delta$ and $u_1 \in \mathcal{S}_n^\delta$ such that

$$\begin{aligned} \Delta_\Psi^\xi \;&:=\; \min\nolimits_{w\in\mathcal{S}_p^\delta} \max\nolimits_{u\in\mathcal{S}_n^\delta} L_\Psi^\xi(w,u) \;-\; \min\nolimits_{w\in\mathcal{S}_w} \max\nolimits_{u\in\mathcal{S}_u} L_\Psi^\xi(w,u) \\ &\leq L_\Psi^\xi(w_1',u_1) - L_\Psi^\xi(w_1,u_1)\;. \end{aligned}$$

Computing the difference gives

$$\begin{aligned} \Delta_\Psi^\xi \;&\leq (w_1'-w_1)^\intercal \mathbf{H} u_1 + \textstyle\sum_{l=1}^M \xi_l\,(\|w_1'\|_{\Sigma^{(l)}} - \|w_1\|_{\Sigma^{(l)}})\|u_1\|_{\tilde{\Sigma}^{(l)}} + (f(w_1',u_1) - f(w_1,u_1)) \\ &\leq \delta\|\mathbf{H}\|K \;+\; S\,K \textstyle\sum_{l=1}^M |\xi_l|\,\|w_1'-w_1\|_{\Sigma^{(l)}} \;+\; |f(w_1',u_1) - f(w_1,u_1)| \\ &\leq \delta K\|\mathbf{H}\| + \delta S^2 K \textstyle\sum_{l=1}^M |\xi_l| + \epsilon\;. \end{aligned}$$

We seek to control $\|\mathbf{H}\|$ and $\sum_{1=1}^M |\xi_l|$ via concentration inequalities. Let vec($\mathbf{H}$) denote the $\mathbb{R}^{pn}$-valued vector formed from the entries of $\mathbf{H}$, and $\Sigma_\mathbf{H} := \text{Var}[\text{vec}(\mathbf{H})]$. Then we can express, for some $\mathbb{R}^{pn}$-valued standard Gaussian vector η,

$$\|\mathbf{H}\|^2 \;=\; \|\text{vec}(\mathbf{H})\|^2 \;=\; \eta^\intercal\, \Sigma_\mathbf{H}\, \eta\;.$$

Then by a Chernoff bound, we have that for any $t > 0$,

$$\mathbb{P}(\|\mathbf{H}\| \geq t) \;\leq\; \inf\nolimits_{a>0} e^{-at^2}\mathbb{E}[e^{a\|\mathbf{H}\|^2}] \;=\; \inf\nolimits_{a>0} e^{-at^2}\mathbb{E}[e^{a\,\eta^\intercal \Sigma_\mathbf{H} \eta}]$$

Applying the formula of the moment-generating function of a Gaussian quadratic form (see e.g. Rencher and Schaalje (2008)) followed by setting $a = \frac{1}{4\|\Sigma_\mathbf{H}\|_{op}}$, we obtain

$$\mathbb{P}(\|\mathbf{H}\| \geq t) \;\leq\; \inf\nolimits_{a>0} \frac{e^{-at^2}}{\sqrt{\det(I_{pn} - 2a\Sigma_\mathbf{H})}} \;\leq\; \frac{e^{-t^2/(4\|\Sigma_\mathbf{H}\|_{op})}}{\sqrt{\det(I_{pn} - \frac{1}{2\|\Sigma_\mathbf{H}\|_{op}}\Sigma_\mathbf{H})}} \;\leq\; 2^{pn/2}\, e^{-t^2/(4\|\Sigma_\mathbf{H}\|_{op})}\;. \tag{104}$$

On the other hand, a standard concentration result on univariate Gaussians yields

$$\mathbb{P}(|\xi_l| > t) \;\leq\; 2e^{-t^2/2}\,.$$

Taking a union bound, we obtain that for any $t > 0$,

$$\mathbb{P}(\,\Delta_\Psi^\xi \;\leq\; \delta K t + \delta S^2 K M t + \epsilon\,) \;\geq\; 1 - 2^{pn/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} - 2Me^{-t^2/2}\,,$$

and therefore for any $c \in \mathbb{R}$ and $t > 0$,

$$\begin{aligned}&\mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_w}\,\max\nolimits_{u\in\mathcal{S}_u}\,L_\Psi^\xi(w,u) \;\geq\; c - \delta K t - \delta S^2 K M t - \epsilon\,)\\ &\quad\geq\; \mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_p^\delta}\,\max\nolimits_{u\in\mathcal{S}_n^\delta}\,L_\Psi^\xi(w,u) \;\geq\; c) - 2^{pn/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} - 2e^{-t^2/2}\,. \end{aligned} \tag{105}$$

A similar argument as in the proof of Theorem 3.2.1 of Thrampoulidis (2016) shows that, there exists $w_2 \in \mathcal{S}_p^\delta$, $u_2 \in \mathcal{S}_w$ and $u_2' \in \mathcal{S}_n^\delta$ with $\|u_2 - u_2'\| \leq \delta$ such that

$$\begin{aligned}\min_{w\in\mathcal{S}_p^\delta}\,\max_{u\in\mathcal{S}_n^\delta}\,L_\psi(w,u) \;-\; \min_{w\in\mathcal{S}_w}\,\max_{u\in\mathcal{S}_u}\,L_\psi(w,u) \;&\geq\; L_\psi(w_2,u_2') - L_\psi(w_2,u_2)\\ &=\; \textstyle\sum_{l=1}^M \Big(\|w_2\|_{\Sigma^{(l)}}\mathbf{h}_l^\intercal(\tilde\Sigma^{(l)})^{1/2}(u_2' - u_2) + w_2^\intercal(\Sigma^{(l)})^{1/2}\mathbf{g}_l(\|u_2'\|_{\tilde\Sigma^{(l)}} - \|u_2\|_{\tilde\Sigma^{(l)}})\Big)\\ &\qquad + (f(w_2,u_2') - f(w_2,u_2))\\ &\geq\; -\delta S^2 K \textstyle\sum_{l=1}^M(\|\mathbf{h}_l\| + \|\mathbf{g}_l\|) - \epsilon\,. \end{aligned}$$

Applying (104) to each $\|\mathbf{h}_l\|$ and $\|\mathbf{g}_l\|$ yields that, for any $t > 0$ and $1 \leq l \leq M$,

$$\mathbb{P}(\|\mathbf{h}_l\| \geq t) \;\leq 2^{n/2}e^{-t^2/4} \qquad\text{and}\qquad \mathbb{P}(\|\mathbf{g}_l\| \geq t) \;\leq 2^{p/2}e^{-t^2/4}\,.$$

Taking another union bound, we get that for any $t > 0$,

$$\begin{aligned}&\mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_w}\,\max\nolimits_{u\in\mathcal{S}_u}\,L_\psi(w,u) \;\geq\; c + 2\delta S^2 K M t + \epsilon\,)\\ &\leq\; \mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_w^\delta}\,\max\nolimits_{u\in\mathcal{S}_u^\delta}\,L_\psi(w,u) \;\geq\; c\,) \;+\; 2^{n/2}Me^{-t^2/4} \;+\; 2^{p/2}Me^{-t^2/4}\,. \end{aligned} \tag{106}$$

Now by Lemma 45, we have

$$\mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_u^\delta}\,\max\nolimits_{u\in\mathcal{S}_d^\delta}\,L_\psi(w,u) \;\geq\; c\,) \;\leq\; \mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_p^\delta}\,\max\nolimits_{u\in\mathcal{S}_n^\delta}\,L_\Psi^\xi(w,u) \;\geq\; c)\,.$$

Combining this with (105) and (106) yields

$$\begin{aligned}&\mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_u}\,\max\nolimits_{u\in\mathcal{S}_d}\,L_\psi(w,u) \;\geq\; c + 2\delta S^2 K M t + \epsilon\,)\\ &\leq\; \mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_u}\,\max\nolimits_{u\in\mathcal{S}_d}\,L_\Psi^\xi(w,u) \;\geq\; c - \delta K t - \delta S^2 K M t - \epsilon\,)\\ &\qquad + 2^{n/2}Me^{-t^2/4} + 2^{p/2}Me^{-t^2/4} + 2^{np/2}\, e^{-t^2/(4\|\Sigma_{\mathbf{H}}\|_{op})} + 2e^{-t^2/2}\,. \end{aligned}$$

The above holds for all $\epsilon > 0$ and $t > 0$. Set $t = \delta^{-1/2}$, take $\epsilon \to 0$ and choosing a sequence $\delta(\epsilon) \to 0$, we obtain that

$$\mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_w}\,\max\nolimits_{u\in\mathcal{S}_u}\,L_\psi(w,u) \;\geq\; c\,) \;\leq\; \mathbb{P}(\,\min\nolimits_{w\in\mathcal{S}_w}\,\max\nolimits_{u\in\mathcal{S}_u}\,L_\Psi^\xi(w,u) \;\geq\; c\,)\,,$$

i.e. $\mathbb{P}(\Psi^\xi_{\mathcal{S}_w,\mathcal{S}_u} \geq c) \;\geq\; \mathbb{P}(\psi_{\mathcal{S}_p,\mathcal{S}_n} \geq c)$. ■

We are now ready to prove Theorem 13 and Corollary 7.

Proof of Theorem 13 The proof is almost identical to the proof of Theorem 3.3.1 of Thrampoulidis (2016) given the GMT result from Lemma 46, and we focus on highlighting the differences. To prove the first bound in (i), we first apply Lemma 46 to obtain that for all $c \in \mathbb{R}$,

$$\mathbb{P}\big(\min_{w \in \mathcal{S}_n} \max_{u \in \mathcal{S}_d} L_\Psi(w,u) + \textstyle\sum_{l=1}^M \xi_l \, \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \leq c \big) \;\leq\; \mathbb{P}(\psi_{S_n, S_d} \leq c) \;,$$

where $(\xi_l)_{l \leq M}$ is a collection of univariate standard Gaussians independent of $\mathbf{H}$. First notice that, by conditioning on the event $\cap_{l \leq M}\{\xi_l \geq 0\}$, we have that

$$\begin{aligned}
\mathbb{P}(\Psi_{\mathcal{S}_p, \mathcal{S}_n} \leq c) \;&= \mathbb{P}\big(\min_{w \in \mathcal{S}_n} \max_{u \in \mathcal{S}_d} L_\Psi(w,u) \leq c \big) \\
&\leq \mathbb{P}\big(\min_{w \in \mathcal{S}_n} \max_{u \in \mathcal{S}_d} L_\Psi(w,u) + \textstyle\sum_{l=1}^M \xi_l \, \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \leq c \,\big|\, \xi_1, \ldots, \xi_M \leq 0 \big)
\end{aligned}$$

which holds almost surely. Since ξ_l's are all independent and symmetric about zero, and there are 2^M possibilities for the signs of $(\xi_1, \ldots, \xi_M)$, we obtain that

$$\begin{aligned}
\frac{1}{2^M} \mathbb{P}(\Psi_{\mathcal{S}_p, \mathcal{S}_n} \leq c) \;&\leq \frac{1}{2^M} \mathbb{P}\big(\min_{w \in \mathcal{S}_n} \max_{u \in \mathcal{S}_d} L_\Psi(w,u) + \textstyle\sum_{l=1}^M \xi_l \, \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \leq c \,\big|\, \xi_1, \ldots, \xi_M \leq 0 \big) \\
&\leq \mathbb{P}\big(\min_{w \in \mathcal{S}_n} \max_{u \in \mathcal{S}_d} L_\Psi(w,u) + \textstyle\sum_{l=1}^M \xi_l \, \|w\|_{\Sigma^{(l)}} \|u\|_{\tilde{\Sigma}^{(l)}} \leq c \big) \\
&\leq \mathbb{P}(\psi_{S_n, S_d} \leq c) \;,
\end{aligned}$$

which gives the desired statement.

The proof of the bound in (ii) is exactly the same as the proof of Theorem 3.3.1(ii) of Thrampoulidis (2016): It relies on the ability to apply a min-max theorem or a min-max inequality for swapping minimum and maximum under the stated convex-concave assumptions, as well as the invariance of the random term of the loss under a sign change. Both hold for our losses L_Ψ and L_ψ, since $\mathbf{H}$ in our L_Ψ is still zero-mean Gaussian, L_ψ is a linear sum of independent mean-zero Gaussian terms and all additional matrices $\Sigma^{(l)}$ and $\tilde{\Sigma}^{(l)}$ are positive semi-definite. We refer readers to the proof of Theorem 3.3.1(ii) of Thrampoulidis (2016) for a detailed derivation, and note that the only difference in our result is in that the coefficient from the first bound in (i) is now 2^M instead of 2.

The proof of (iii) is also exactly the same as the proof of Theorem 3.3.1(iii) of Thrampoulidis (2016), which only relies on the three assumptions, the statements (i) and (ii) proved above and a union bound. We again refer readers to the proof of Theorem 3.3.1(iii) of Thrampoulidis (2016) for a detailed derivation. ■

Proof of Corollary 7 The result follows directly from Theorem 13(ii); see Corollary 3.3.2 of Thrampoulidis (2016).

■

Appendix M. Intermediate results for applying CGMT to data augmentation

For clarity, throughout Appendices M and N, we will index all augmentations as ϕ_{ij} where $i \leq m$, the number of original data, and $j \leq k$, the number of augmentations. Recall that $n = mk$. We also write the label of $\phi_{ij}(Z_i)$ as $y_i(Z_i)$ to emphasize the dependence on the original data Z_i.

M.1. Equivalence of different optimization problems

To prove Theorem 12, we seek to obtain a set of deterministic equations whose solutions characterize the high-dimensional behavior of logistic regression estimate. This involves establishing the equivalence of a series of optimization problems, which are defined in this section. We also formally state all lemmas used to establish the equivalence.

Original optimization (OO). The loss on the augmented data computed on $\beta \in \mathbb{R}^p$ is given as

$$L_\beta(\mathbf{X},\mathbf{X}^\Phi) \;:=\; \frac{1}{mk}\sum\nolimits_{i=1}^m \sum\nolimits_{j=1}^k \Big(\log\big(1+e^{(\phi_{ij}(Z_i))^\intercal\beta}\big) - y_i(Z_i)\times(\phi_{ij}(Z_i))^\intercal\beta\Big) + \frac{\lambda}{2n}\|\beta\|_2^2 \;. \tag{107}$$

Here, the loss is computed on the two dependent $\mathbb{R}^{m\times p}$ and $\mathbb{R}^{mk\times p}$-valued data matrices

$$\mathbf{X} \;:=\; \begin{pmatrix} \leftarrow Z_1^\intercal \rightarrow \\ \vdots \\ \leftarrow Z_m^\intercal \rightarrow \end{pmatrix} \quad\text{and}\quad \mathbf{X}^\Phi \;:=\; \begin{pmatrix} \mathbf{X}_1^\Phi \\ \vdots \\ \mathbf{X}_m^\Phi \end{pmatrix} \quad\text{where}\quad \mathbf{X}_i^\Phi \;:=\; \begin{pmatrix} \leftarrow(\phi_{i1}(Z_i))^\intercal\rightarrow \\ \vdots \\ \leftarrow(\phi_{ik}(Z_i))^\intercal\rightarrow \end{pmatrix}.$$

Let S be any convex and compact subset of $\mathbb{R}^p$. We denote the minimized risk over S and the corresponding minimizer respectively as

$$\hat{R}_S(\mathbf{X},\mathbf{X}^\Phi) \;:=\; \min_{\beta\in S} L_\beta(\mathbf{X},\mathbf{X}^\Phi) \qquad\text{and}\qquad \hat{\beta}_S(\mathbf{X},\mathbf{X}^\Phi) \;:=\; \operatorname*{arg\,min}_{\beta\in S} L_\beta(\mathbf{X},\mathbf{X}^\Phi)\;. \tag{OO}$$

We label (OO) as the **original optimization**. By our universality result, we may replace the dependent data matrices in (OO) by Gaussian matrices.

Gaussian optimization (GO). Recall that $\Sigma_o = \mathrm{Var}[Z_1]$ and $\Sigma = \mathrm{Var}[\phi_{11}(Z_1)]$. We denote the corresponding minimized risk under Gaussian data as

$$\begin{aligned} &\hat{R}_S(\mathbf{G}\Sigma_o^{1/2},\mathbf{G}^\Phi\Sigma^{1/2}) \;:=\; \min\nolimits_{\beta\in S} L_\beta(\mathbf{G}\Sigma_o^{1/2},\mathbf{G}^\Phi\Sigma^{1/2})\;,\\ \text{and}\qquad &\hat{\beta}_S(\mathbf{G}\Sigma^{1/2},\mathbf{G}^\Phi\Sigma_o^{1/2}) \;:=\; \operatorname*{arg\,min}_{\beta\in S} L_\beta(\mathbf{G}\Sigma^{1/2},\mathbf{G}^\Phi\Sigma_o^{1/2})\;. \end{aligned} \tag{GO}$$

The risk is computed on the two correlated Gaussian matrices

$$\mathbf{G} \;:=\; \begin{pmatrix} \leftarrow G_1^\intercal \rightarrow \\ \vdots \\ \leftarrow G_m^\intercal \rightarrow \end{pmatrix} \quad\text{and}\quad \mathbf{G}^\Phi \;:=\; \begin{pmatrix} \mathbf{G}_1^\Phi \\ \vdots \\ \mathbf{G}_m^\Phi \end{pmatrix} \quad\text{where}\quad \mathbf{G}_i^\Phi \;:=\; \begin{pmatrix} \leftarrow(G_{i1}^\Phi)^\intercal\rightarrow \\ \vdots \\ \leftarrow(G_{ijk}^\Phi)^\intercal\rightarrow \end{pmatrix},$$

where $\Sigma_o^{1/2}G_i$ corresponds to Z_i, $\mathbf{G}_i^\Phi\Sigma^{1/2}$ corresponds to $\mathbf{X}_i^\Phi$ and $\Sigma^{1/2}G_{ij}^\Phi$ corresponds to $\phi_{ij}(Z_i)$, and

$$\mathbb{E}[(\mathbf{G},\mathbf{G}^\Phi)] \;=\mathbb{E}[(\mathbf{X},\mathbf{X}^\Phi)] \;=\; 0 \qquad\text{and}\qquad \mathrm{Var}[(\mathbf{G}\Sigma_o^{1/2},\mathbf{G}^\Phi\Sigma^{1/2})] \;=\mathbb{E}[(\mathbf{X},\mathbf{X}^\Phi)]\;.$$

Primary optimization (PO). Since (GO) only depends on Gaussian data, we may adapt the CGMT technique to analyze its limiting behaviour. This requires a reformulation of (GO) in a similar way to the reformulation of the primary optimization in Salehi et al. (2019). To make this reformulation precise, we introduce some more notations. Given an $\mathbb{R}^{mk}$-valued vector $\boldsymbol{v}$, we denote

$$\rho(\boldsymbol{v}) \;:=\; \big(\log(1+e^{v_{11}}),\,\ldots\,,\,\log(1+e^{v_{mk}})\big)^\intercal \;\in\; \mathbb{R}^{mk}\;.$$

Also write the $\mathbb{R}^{mk}$-valued vector of labels for (the Gaussian surrogates for) the augmented data as

$$\mathbf{y}(\mathbf{G}\Sigma_o^{1/2}\beta^*) \;\coloneqq\; (\underbrace{y_1(\Sigma_o^{1/2}G_1), \ldots, y_1(\Sigma_o^{1/2}G_1)}_{\text{repeated } k \text{ times}}, \;\ldots,\; \underbrace{y_m(\Sigma_o^{1/2}G_m), \ldots, y_m(\Sigma_o^{1/2}G_m)}_{\text{repeated } k \text{ times}})^\intercal \;,$$

where we highlight that $\mathbf{y}$ depends on $\mathbf{G}$ only through the $\mathbb{R}^n$ vector $\mathbf{G}\Sigma_o^{1/2}\beta^*$. For $d \in \mathbb{N}$, we also write $\mathbf{1}_d$ as the all-one vector in $\mathbb{R}^d$.This allows us to rewrite the loss in (GO) as

$$L_\beta(\mathbf{G}\Sigma_o^{1/2}, \mathbf{G}^\Phi\Sigma^{1/2}) \;=\; \frac{1}{mk}\mathbf{1}_{mk}^\intercal\rho(\mathbf{G}^\Phi\Sigma^{1/2}\beta) - \frac{1}{mk}\mathbf{y}(\mathbf{G}\Sigma_o^{1/2}\beta^*)^\intercal\mathbf{G}^\Phi\Sigma^{1/2}\beta + \frac{\lambda}{2n}\|\beta\|_2^2 \;.$$

Introducing a new variable $u \in \mathbb{R}^{mk}$ and a corresponding Lagrange multiplier $v \in \mathbb{R}^{mk}$, we can consider an alternative loss

$$L_{\beta,u,v}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;\coloneqq\; \frac{1}{mk}\mathbf{1}_{mk}^\intercal\rho(u) - \frac{1}{mk}\mathbf{y}(\mathbf{G}\Sigma_o^{1/2}\beta^*)^\intercal u + \frac{\lambda}{2n}\|\beta\|_2^2 + \frac{1}{mk}v^\intercal(u - \mathbf{G}^\Phi\Sigma^{1/2}\beta) \;.$$

For subsets $S \subseteq \mathbb{R}^p$ and $S_u, S_v \subseteq \mathbb{R}^{mk}$, we denote the minimized loss and the minimizer as

$$\begin{aligned} &R_{S,S_u,S_v}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;\coloneqq\; \min_{\beta\in S, u\in S_u}\;\max_{v\in S_v} L_{\beta,u,v}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \\ \text{and}\quad &\beta_{S,S_u,S_v}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;\coloneqq\; \arg\min_{\beta\in S}\;\min_{u\in S_u}\;\max_{v\in S_v} L_{\beta,u,v}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;. \end{aligned} \tag{PO}$$

Lemma 47 (Equivalence of (GO) and (PO))

$$\begin{aligned} \hat{R}_S(\mathbf{G}\Sigma_o^{1/2}, \mathbf{G}^\Phi\Sigma^{1/2}) \;&= R_{S,\mathbb{R}^{mk},\mathbb{R}^{mk}}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;, \\ \hat{\beta}_S(\mathbf{G}\Sigma_o^{1/2}, \mathbf{G}^\Phi\Sigma^{1/2}) \;&= \beta_{S,\mathbb{R}^{mk},\mathbb{R}^{mk}}^{\mathrm{PO}}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2}) \;. \end{aligned}$$

Proof of Lemma 47 The proof is exactly the same to the reformulation of the primary optimization in Salehi et al. (2019) by the Lagrange multiplier method, and we refer readers to their (37) – (40) in Appendix C for the proof. ■

Auxiliary optimization (AO). Before we present the auxiliary optimization, we notice that two key issues make our problem more complicated from the setup in Salehi et al. (2019):

- In Salehi et al. (2019), they have the same data matrices for $\mathbf{G}$ and $\mathbf{G}^\Phi$ with i.i.d. standard normal entries and $\Sigma_o = \Sigma = I_d$. This allows them to project $\mathbf{G}^\Phi$ onto the subspace orthogonal to β^*, which is independent of $\mathbf{G}\beta^*$, and apply CGMT. In our case, $\mathbf{G}$ and $\mathbf{G}^\Phi$ are different and have non-trivial dependence. We instead make use of a projection $P_*^\perp$ adapted to the variance-covariance structure in Assumption 11, defined through

$$P_* \;\coloneqq\; \begin{cases} \frac{(\Sigma_*\Sigma_o^{1/2}\beta^*)(\Sigma_*\Sigma_o^{1/2}\beta^*)^\intercal}{\|\Sigma_*\Sigma_o^{1/2}\beta^*\|^2} & \text{if } \Sigma_*\Sigma_o^{1/2}\beta^* \neq 0 \\ 0 & \text{otherwise} \;, \end{cases} \qquad \text{and} \qquad P_*^\perp \;\coloneqq I_p - P_* \;.$$

  In other words, $P_*^\perp$ is a projection onto the subspace orthogonal to $\Sigma_*\Sigma_o^{1/2}\beta^*$. This is explicitly addressed in Appendix N.1;

- As discussed in Section 5, the Gaussian matrix handled by existing work on CGMT is either one with i.i.d. coordinates, or one formed by multiplying a coordinate-wise i.i.d. matrix by an $\mathbb{R}^{m\times m}$ matrix and an $\mathbb{R}^{p\times p}$ matrix from both sides. Our augmented matrix, $\mathbf{G}^\Phi$, cannot be expressed in either form due to the simultaneous presence of two forms of variances: Each row of $\mathbf{G}^\Phi$ admits a variance of I_p, whereas the rows corresponding to different augmentations of the same data admit a variance of Σ_*. We resolve this issue by applying our dependent CGMT (Theorem 5) with $M = 2$.

Having addressed these two issues, we are able to borrow most of the algebraic calculations from Salehi et al. (2019) for analyzing the auxiliary optimization, except that the limiting terms we obtain are different due to augmentations.

To state the auxiliary optimization, let $\mathbf{g}_1, \mathbf{g}_2, \mathbf{h}_1, \mathbf{h}_2$ be independent standard Gaussians such that $\mathbf{g}_l$'s are $\mathbb{R}^p$-valued and $\mathbf{h}_l$'s are $\mathbb{R}^{mk}$-valued. We also denote the collections $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2)$ and $\mathbf{h} = (\mathbf{h}_1, \mathbf{h}_2)$ for short, and define the matrices

$$\Sigma_1 \;:=\; \Sigma^{1/2} P_*^\perp (I_p - \Sigma_*) P_*^\perp \Sigma^{1/2} \;, \qquad \Sigma_2 \;:=\; \Sigma^{1/2} P_*^\perp \Sigma_* P_*^\perp \Sigma^{1/2} \;, \qquad J_{mk} \;:=\; \begin{pmatrix} \mathbf{1}_{k\times k} & & \\ & \ddots & \\ & & \mathbf{1}_{k\times k} \end{pmatrix} \in \mathbb{R}^{mk\times mk} \;.$$

The loss of (AO), parameterized by $\beta \in \mathbb{R}^p$ and $u, v \in \mathbb{R}^{mk}$, is given as

$$\begin{aligned} L^{\mathrm{AO}}_{\beta,u,v}(\mathbf{y}, \mathbf{G}^\Phi P_*, \mathbf{g}, \mathbf{h}) \;:=\; & \frac{1}{mk} \mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk} \mathbf{y}^\top u + \frac{\lambda}{2n} \|\beta\|_2^2 + \frac{1}{mk} v^\top (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta) - \frac{1}{mk} v^\top \mathbf{h}_1 \|\beta\|_{\Sigma_1} \\ & - \frac{1}{mk} \|v\| \mathbf{g}_1^\top \Sigma_1^{1/2} \beta - \frac{1}{mk^{3/2}} v^\top J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2} - \frac{1}{mk} \|v\|_{J_{mk}} \mathbf{g}_2^\top \Sigma_2^{1/2} \beta \;. \end{aligned}$$

Note that we have abbreviated $\mathbf{y} = \mathbf{y}(\mathbf{G}\Sigma_o^{1/2}\beta^*)$. We also denote the minimized loss with respect to the subset $(S, S_u, S_v) \subseteq \mathbb{R}^p \times \mathbb{R}^{mk} \times \mathbb{R}^{mk}$ as

$$R^{\mathrm{AO}}_{S,S_u,S_v}(\mathbf{y}, \mathbf{G}^\Phi P_*, \mathbf{g}, \mathbf{h}) \;:=\; \min_{\beta \in S, u \in S_u} \max_{v \in S_v} L^{\mathrm{AO}}_{\beta,u,v}(\mathbf{y}, \mathbf{G}^\Phi P_*, \mathbf{g}, \mathbf{h}) \;. \tag{AO}$$

The next result applies Theorem 5 to convert (PO) into (AO).

Lemma 48 (Equivalence of (PO) and (AO)) *Suppose Assumption 11 holds. Let $S \subset \mathbb{R}^p$ and $S_u, S_v \in \mathbb{R}^{mk}$ be compact, convex and non-empty. Then all conclusions of Theorem 5 hold with $\Psi_{\mathcal{S}_p,\mathcal{S}_n}$ replaced by $R^{\mathrm{PO}}_{S,S_u,S_v}(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi\Sigma^{1/2})$ and $\psi_{\mathcal{S}_p,\mathcal{S}_m}$ replaced by $R^{\mathrm{AO}}_{S,S_u,S_v}(\mathbf{y}, \mathbf{G}^\Phi P_*, \mathbf{g}, \mathbf{h})$.*

Scalar optimization (SO). The next step is to convert (AO) into a scalar formulation. For convenience we write $\|\bullet\| = \|\bullet\|_2$ as the Euclidean norm throughout this section, unless otherwise specified. While the form of the optimization is complicated, we note that the terms are largely similar to the AO in Salehi et al. (2019), except for additional parameters $(\sigma_1, \nu_1, r_1, \tau_1)$ introduced to handle the additional covariance across different augmented versions of the same data. To define the scalar formulation, given the convex compact and non-empty subsets $S \subset \mathbb{R}^p$, $S_u, S_v \in \mathbb{R}^{mk}$, we define the following compact domains of optimization:

$$S_{r_1} \;:=\; \Big\{ \frac{1}{\sqrt{mk}} \|P_{mk}^\perp v\| \,\Big|\, v \in S_v \Big\} \;, \qquad S_{r_2} \;:=\; \Big\{ \frac{1}{\sqrt{mk}} \|P_{mk} v\| \,\Big|\, v \in S_v \Big\} \;,$$

where we have defined the projection matrices $P_{mk} \coloneqq \frac{1}{k} J_{mk}$ and write $P^{\perp}_{mk} = I_{mk} - P_{mk}$. Also define

$$S^{\alpha} \;\coloneqq\; \left\{ \frac{v(\beta^*)^{\intercal}\Sigma^{1/2}\beta}{\sqrt{p}\,\kappa_*^2} \,\middle|\, \beta \in S \right\} ,$$

where $v(\beta^*) \coloneqq \sqrt{p}\,\Sigma_*\Sigma_o^{1/2}\beta^*$, $\kappa_* = \|\Sigma_*\Sigma_o^{1/2}\beta^*\|$. Define

$$S_{\sigma_1} \;\coloneqq\; \left\{ \|(I_p - \Sigma_*)P_*^{\perp}\Sigma^{1/2}\beta\| \,\middle|\, \beta \in S \right\} , \qquad S_{\sigma_2} \;\coloneqq\; \left\{ \|\Sigma_* P_*^{\perp}\Sigma^{1/2}\beta\| \,\middle|\, \beta \in S \right\} .$$

The optimization will be performed over the two $\mathbb{R}^5$-valued vectors

$$\begin{aligned} \tilde{\alpha} \;&\coloneqq (\alpha, \sigma_1, \sigma_2, v_1, v_2) \;\in\; S^{\alpha} \times S_{\sigma_1} \times S_{\sigma_2} \times (\mathbb{R}_0^+)^2 \;\coloneqq\; S_1 , \\ \tilde{\theta} \;&\coloneqq (r_1, r_2, \tau_1, \tau_2, \theta) \;\in\; S_{r_1} \times S_{r_2} \times (\mathbb{R}_0^+)^2 \times \mathbb{R} \;\coloneqq\; S_2 . \end{aligned}$$

We also define $P_{\Sigma} = (\Sigma^{\dagger})^{1/2}\Sigma^{1/2}$, the projection onto the positive eigenspace of Σ, and the matrix

$$\tilde{\Sigma}_{\sigma,\tau} \;\coloneqq\; \frac{1}{2\sigma_1\tau_1}(P_{\Sigma} - \Sigma_*) + \frac{1}{2\sigma_2\tau_2}\Sigma_* \;.$$

Also define the Gaussian random vectors

$$\begin{aligned} \mathbf{q} \;&\coloneqq\; \frac{1}{\kappa_*\;\sqrt{p}}\mathbf{G}^{\Phi}\, v(\beta^*) \;=\; \mathbf{G}^{\Phi}\,\frac{\Sigma_*\Sigma_o^{1/2}\beta^*}{\|\Sigma_*\Sigma_o^{1/2}\beta^*\|} \;, \qquad \tilde{\mathbf{h}}_{\alpha,\sigma} \;\coloneqq\; \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}}J_{mk}\mathbf{h}_2 \;, \\ \tilde{\mathbf{g}} \;&\coloneqq\; -\frac{r_1 + r_2}{\sqrt{mk}}\,(P_{\Sigma} - \Sigma_*)\mathbf{g}_1 - \frac{r_2}{\sqrt{m}}\,\Sigma_*\mathbf{g}_2 \;. \end{aligned}$$

For a function $f : \mathcal{S}' \to \mathbb{R}$ and some $\mathcal{S}' \subseteq \mathbb{R}^{mk}$, we define the Moreau envelope

$$\mathcal{M}_{\mathcal{S}}(f; v, t) \;\coloneqq\; \min_{x\in\mathcal{S}} f(x) + \frac{1}{2t}\|x - v\|_2^2 \;.$$

Now we are ready to define the loss

$$\begin{aligned} L^{\mathrm{SO}}_{\tilde{\alpha},\tilde{\theta}}(\mathbf{y}, \mathbf{q}, \mathbf{g}, \mathbf{h}) \;\coloneqq\; &-\frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2v_1} + \frac{r_2}{2v_2} + \alpha\theta\kappa_*^2 - \frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2} + M_{\mathbf{g},\sigma,\tau,\theta} - \frac{1}{4}\Big\|(\tilde{\Sigma}^{\dagger}_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2 \\ &+ \frac{1}{mk}M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,v} - \frac{1}{2r_2v_2mk}\|\mathbf{y}\|^2 - \frac{1}{mk}\mathbf{y}^{\intercal}\tilde{\mathbf{h}}_{\alpha,\sigma} \;, \end{aligned}$$

where we have defined the nested Moreau envelope $M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,v}$ via

$$\begin{aligned} M^{\perp}_{\tilde{\mathbf{h}}_{\alpha,\sigma},r,v}(\tilde{u}) \;&\coloneqq \mathcal{M}_{P^{\perp}_{mk}(S_u)}(\mathbf{1}^{\intercal}_{mk}\,\rho(\tilde{u} + \bullet)\,;\, P^{\perp}_{mk}\tilde{\mathbf{h}}_{\alpha,\sigma}, \frac{1}{r_1v_1}) \;, \\ M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,v} \;&\coloneqq \mathcal{M}_{P_{mk}(S_u)}(M^{\perp}_{\tilde{\mathbf{h}}_{\alpha,\sigma},r,v}\,;\, \frac{1}{r_2v_2}\mathbf{y} - P_{mk}\tilde{\mathbf{h}}_{\alpha,\sigma}\,,\, r_2v_2) \;, \end{aligned}$$

as well as another Moreau envelope like term

$$M_{\mathbf{g},\sigma,\tau,\theta} \;\coloneqq\; \min_{\mu\in S}\; \frac{\lambda}{2n}\|P_{\Sigma}\mu\|_2^2 + \Big\|\tilde{\Sigma}^{1/2}_{\sigma,\tau}(\Sigma^{1/2}\mu) - \frac{1}{2}(\tilde{\Sigma}^{\dagger}_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2 + \frac{r_2}{\sqrt{n}}\mathbf{g}_2^{\intercal}P_*\Sigma^{1/2}\mu \;.$$

The minimized risk is denoted as

$$R^{\mathrm{SO}}_{S,S_u,S_v}(\mathbf{y}, \mathbf{q}, \mathbf{g}, \mathbf{h}) \;\coloneqq\; \min_{\tilde{\alpha}\in S_1}\;\max_{\tilde{\theta}\in S_2}\;\min_{\tilde{\chi}\in S_3} L^{\mathrm{SO}}_{\tilde{\alpha}\tilde{\theta}}(\mathbf{y}, \mathbf{q}, \mathbf{g}, \mathbf{h}) \;. \tag{SO}$$

The next lemma shows that (AO) can be replaced by (SO) in that it satisfies similar inequalities as (AO) in terms of their relationships to (PO). The inequalities in the result are to be compared with those in Theorem 5.

Lemma 49 (Equivalence of (PO) and (SO)) *Let $S \in \mathbb{R}^p$ and $S_u, S_v \in \mathbb{R}^{mk}$ be compact, convex and non-empty. Also assume that the linear span span$(S) = \mathbb{R}^p$. Then for any $c \in \mathbb{R}$,*

$$\mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v} \leq c) \;\leq 4\,\mathbb{P}(R^{\mathrm{SO}}_{S,S_u,S_v} \;\leq\; c) \qquad \textit{and} \qquad \mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v} \geq c) \;\leq 4\,\mathbb{P}(R^{\mathrm{SO}}_{S,S_u,S_v} \;\geq\; c)\;.$$

If instead of S, the set of values of β we consider is the non-convex set

$$S_{c,\epsilon} \;:=\; S \setminus \{\beta \in S \;\big|\; |(P_\Sigma\beta)^\intercal \Sigma_o (P_\Sigma\beta) - c| \leq \epsilon\;\}$$

for some $c \in \mathbb{R}$ and some sufficiently small $\epsilon > 0$ such that $S_{c,\epsilon}$ is non-empty. Then we have

$$\mathbb{P}(R^{\mathrm{PO}}_{S_{c,\epsilon},S_u,S_v} \leq c) \;\leq 4\,\mathbb{P}(R^{\mathrm{SO}}_{S_{c,\epsilon},S_u,S_v} \;\leq\; c)\;.$$

Deterministic optimization (DO). The next step is to compute the asymptotics of (SO) as $m, p \to \infty$ and $p/m \to \kappa/k$ for special cases of $S \subset \mathbb{R}^p$. The limit is given by a deterministic optimization

$$\begin{aligned} R^{DO}_S \;:=\; \min_{\substack{\alpha \in S^\alpha \\ (\sigma_1,\sigma_2)\in S_{\sigma_1}\times S_{\sigma_2} \\ \nu_1,\nu_2 \geq 0}} \;\max_{\substack{(r_1,r_2)\in S_{r_1}\times S_{r_2} \\ \tau_1,\tau_2\geq 0 \\ \theta\in\mathbb{R}}} \;& -\frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} + \alpha\theta\bar\kappa_*^2 - \frac{\alpha^2\bar\kappa_*^2}{2\sigma_2\tau_2} - \bar\chi_1^{r,\theta,\sigma,\tau} + \epsilon_S^2\,\frac{\bar\chi_3^{r,\theta,\sigma,\tau}}{\bar\chi_2^{r,\theta,\sigma,\tau}} \\ & - \frac{1}{4r_2\nu_2} - \alpha\,\mathbb{E}[\sigma(\bar\kappa_o\bar Z_0 + \bar\kappa_*\bar Z_1)\bar\kappa_*\bar Z_1] + \bar M_\rho^{r,\nu,\alpha,\sigma}\;, \end{aligned} \tag{DO}$$

where we have defined the limits

$$\begin{aligned} \bar\kappa_* \;&:=\; \lim_{p\to\infty}\kappa_* \;=\; \lim_{p\to\infty} \|\Sigma_*\Sigma_o^{1/2}\beta^*\|\;, \qquad \bar\kappa_o \;:=\; \lim_{p\to\infty}\|(I_p - \Sigma_*)\Sigma_o^{1/2}\beta^*\|\;, \\ \bar\chi_1^{r,\theta,\sigma,\tau} \;&:=\; \frac{(r_1+r_2)^2\sigma_1\tau_1}{2k}\,\bar\chi_{11}^{\sigma,\tau} + \frac{r_2^2\sigma_2\tau_2}{2}\,\bar\chi_{12}^{\sigma,\tau} + \frac{\theta^2\bar\kappa_*^2\sigma_2\tau_2}{2}\,\bar\chi_{13}^{\sigma,\tau}\;, \\ \bar\chi_2^{r,\theta,\sigma,\tau} \;&:=\; \frac{(r_1+r_2)^2\sigma_1^2\tau_1^2}{k}\bar\chi_{21}^{\sigma,\tau} + r_2^2\sigma_2^2\tau_2^2\,\bar\chi_{22}^{\sigma,\tau} + \theta^2\bar\kappa_*^2\sigma_2^2\tau_2^2\,\bar\chi_{23}^{\sigma,\tau}\;, \\ \bar\chi_3^{r,\theta,\sigma,\tau} \;&:=\; \frac{(r_1+r_2)^2\sigma_1^2\tau_1^2}{k}\bar\chi_{31}^{\sigma,\tau} + r_2^2\sigma_2^2\tau_2^2\,\bar\chi_{32}^{\sigma,\tau} + \theta^2\bar\kappa_*^2\sigma_2^2\tau_2^2\,\bar\chi_{33}^{\sigma,\tau}\;, \end{aligned}$$

with

$$\begin{aligned} \bar\chi_{11}^{\sigma,\tau} \;&:=\; \lim \frac{\mathrm{Tr}\,((\frac{\sigma_1\tau_1\lambda}{m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*))}{m}\;, \\ \bar\chi_{12}^{\sigma,\tau} \;&:=\; \lim \frac{\mathrm{Tr}\,((\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p)^\dagger\Sigma_*)}{m}\;, \\ \bar\chi_{13}^{\sigma,\tau} \;&:=\; \lim \mathrm{Tr}\Big(\Big(\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big)\;, \\ \bar\chi_{21}^{\sigma,\tau} \;&:=\; \lim \frac{\big\|\Sigma_{\mathrm{new}}^{1/2}(\Sigma^\dagger)^{1/2}(\frac{\sigma_1\tau_1\lambda}{m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*)\big\|^2}{m}\;, \\ \bar\chi_{22}^{\sigma,\tau} \;&:=\; \lim \frac{\big\|\Sigma_{\mathrm{new}}^{1/2}(\Sigma^\dagger)^{1/2}\big(\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p\big)^\dagger\Sigma_*\big\|^2}{m}\;, \\ \bar\chi_{23}^{\sigma,\tau} \;&:=\; \lim \Big\|\Sigma_{\mathrm{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big\|^2\;, \\ \bar\chi_{31}^{\sigma,\tau} \;&:=\; \lim \frac{\big\|(\frac{\lambda}{2m}\Sigma^\dagger + \tilde\Sigma_{\sigma,\tau})^{1/2}\Sigma^{1/2}P_{\Sigma_{\mathrm{new}}}(\Sigma^\dagger)^{1/2}(\frac{\sigma_1\tau_1\lambda}{m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*)\big\|^2}{m}\;, \\ \bar\chi_{32}^{\sigma,\tau} \;&:=\; \lim \frac{\big\|(\frac{\lambda}{2m}\Sigma^\dagger + \tilde\Sigma_{\sigma,\tau})^{1/2}\Sigma^{1/2}P_{\Sigma_{\mathrm{new}}}(\Sigma^\dagger)^{1/2}\big(\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p\big)^\dagger\Sigma_*\big\|^2}{m}\;, \\ \bar\chi_{33}^{\sigma,\tau} \;&:=\; \lim \Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde\Sigma_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}P_{\Sigma_{\mathrm{new}}}(\Sigma^\dagger)^{1/2}\Big(\frac{\sigma_2\tau_2\lambda}{m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big\|^2\;. \end{aligned}$$

We have also defined an expected Moreau-envelope-like term

$$\bar{M}_\rho^{r,\nu,\alpha,\sigma} := \mathbb{E}\Big[\min_{\tilde{u}\in\mathbb{R}^k} \frac{1}{k}\mathbf{1}_k^\intercal \rho(\tilde{u}) + \frac{r_1\nu_1}{2k} \Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(\tilde{u} + \sigma_1\eta)\Big\|^2 \\ + \frac{r_2\nu_2}{2k} \Big\| \frac{1}{k}\mathbf{1}_{k\times k}\Big(\tilde{u} - \frac{1}{r_2\nu_2}\mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\Big\|^2 \Big],$$

where $\bar{Z}_0, \bar{Z}_1, \bar{Z}_2, \eta_1, \ldots, \eta_k$ are i.i.d. univariate Gaussians and $\eta = (\eta_1, \ldots, \eta_k)$, and ε_1 is an independent Logistic(0, 1) variable. The two cases of S we consider are

$$S = \mathcal{S}_p \qquad \text{and} \qquad S = \mathcal{S}_\epsilon^c := \{\beta \in \mathcal{S}_p \,\big|\, \big|\sqrt{\beta^\intercal \Sigma_{\text{new}}\beta} - (\bar{\chi}_2^{\bar{r},\bar{\theta},\bar{\sigma},\bar{\tau}})^{1/2}\big| > \epsilon\},$$

where $\bar{r} = (\bar{r}_1, \bar{r}_2)$, $\bar{\sigma} = (\bar{\sigma}_1, \bar{\sigma}_2)$, $\bar{\theta}$ and $\bar{\tau} = (\bar{\tau}_1, \bar{\tau}_2)$ are the optimal solutions to (DO). We also set $\epsilon_S = 0$ for $S = \mathcal{S}_p$ and $\epsilon_S = \epsilon$ for $S = \mathcal{S}_\epsilon^c$.

Lemma 50 (Equivalence between (SO) and (DO)) *Assume that the set $S_u \subset \mathbb{R}^{mk}$ is closed under permutation of the m blocks of k coordinates. Also suppose that as $m, p \to \infty$, $\sup_{u\in S_u} \frac{\|u\|_2^2}{mk} \to \infty$ and $\sup_{u\in S_v} \frac{\|v\|_2^2}{mk} \to \infty$. Also assume that the limits $\bar{\kappa}_*$, $\bar{\chi}_1^{r,\theta,\sigma,\tau}$, $\bar{\chi}_2^{r,\theta,\sigma,\tau}$ and $\bar{\chi}_3^{r,\theta,\sigma,\tau}$ exist for every $r_1, r_2, \theta, \sigma_1, \sigma_2, \tau_1, \tau_2$. Then for $S = \mathcal{S}_p$ and $S = \mathcal{S}_\epsilon^c$,*

$$\big| R^{\text{SO}}_{S,S_u,S_v}(\mathbf{y}, \mathbf{q}, \mathbf{g}, \mathbf{h}) - R^{\text{DO}}_S \big| \xrightarrow{\mathbb{P}} 0\,.$$

As with Salehi et al. (2019), it remains to prove that the first order condition of (DO) for $S = \mathcal{S}_p$ is equivalent to the system of 10 equations (EQs) in $(\alpha, \sigma_1, \sigma_2, \tau_1, \tau_2, \nu_1, \nu_2, r_1, r_2, \theta)$. This involves computing the derivative of the Moreau-envelope-like term $\bar{M}_\rho^{r,\nu,\alpha,\sigma}$ using the envelope theorem.

Lemma 51 *Assume that the minimizer-maximizers of (DO) are within the interior of the domain of optimization and that $S = \mathcal{S}_p$. Then these minimizer-maximizers are solutions to (EQs).*

M.2. Verifying conditions for different augmnetations

Isotropic data with no augmentation. Salehi et al. (2019) derives a set of equations that governs the behavior of high-dimensional logistic regression with ridge regularization, isotropic data and no data augmentation. Here, we verify that our formula recover their formula exactly as a special case, and that $(r_2, \nu_2, \sigma_2, \tau_2, \alpha, \theta)$ play the role of the parameters in the original unaugmented optimization.

Lemma 52 *Suppose that $X_{\text{new}} \overset{d}{=} Z_1$ with $\text{Var}[Z_1] = \frac{1}{p}I_p$, that $k = 1$ and $\phi_1(Z_i) = Z_i$ almost surely for all $i \leq m = n$. Also write $\gamma = \frac{1}{r_2\nu_2}$, $\rho(\,\bullet\,) = \log(1 + \exp(\,\bullet\,))$ and denote the proximal operator $\text{Prox}_{t\rho(\,\bullet\,)}(v) := \arg\min_{x\in\mathbb{R}} \frac{1}{2t}(v - x)^2 + \rho(x)$. Then (EQs) is equivalent to the following system of* *equations:*

$$\begin{cases} \theta \;=\; \dfrac{\alpha\kappa}{\gamma} \,, \\ \tau_2 \;=\; \dfrac{\kappa^{-1}\gamma}{\sigma_2(1-\gamma\lambda)} \,, \\ r_2 \;=\; \dfrac{\sigma_2\sqrt{\kappa}}{\gamma} \,, \\ \dfrac{\sigma^2\kappa}{2} \;=\; \mathbb{E}[\partial\rho(-\bar{\kappa}_*\bar{Z}_1)(\alpha\bar{\kappa}_*\bar{Z}_1+\sigma_2\bar{Z}_2-\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1+\sigma_2\bar{Z}_2))^2] \,, \\ 1-\kappa+\gamma\lambda\kappa \;=\; \mathbb{E}\Big[\dfrac{2\rho'(-\bar{\kappa}_*Z_1)}{1+\gamma\rho''(\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\bar{\kappa}_*\alpha\bar{Z}_1+\sigma_2\bar{Z}_2))}\Big] \,, \\ -\dfrac{\alpha\kappa}{2} \;=\; \mathbb{E}[\partial^2\rho(-\bar{\kappa}_*\bar{Z}_1)\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\bar{\kappa}_*\alpha\bar{Z}_1+\sigma_2\bar{Z}_2)] \,, \end{cases}$$

with $r_1 = \sigma_1 = 0$, $\nu_1, \tau_1 \to \infty$, $\bar{\kappa}_* = \lim_{p\to\infty} \frac{\|\beta^*\|}{\sqrt{p}}$ and $\kappa = \lim p/n$.

Remark 53 *(i) Lemma 51 and Theorem 12 apply even though the values of r_1, σ_1, ν_1 and τ_1 are not in the interior of the domain of optimization, as these variables can be removed much earlier in the proof of Lemma 49 and allow us to handle only a smaller system of equations. Moreover, the only quantity $\bar{\chi}_2^{r,\theta,\sigma,\tau}$ that enters the test risk is independent of these variables. (ii) To identify the equations in Lemma 52 with those in (14) and (16) from Theorem 2 of Salehi et al. (2019), we note several notational differences: We have used $\kappa = \lim p/n$, whereas they use $\delta = \lim n/p$; our $\bar{\kappa}_*$, $\bar{Z}_1$ and $\bar{Z}_2$ should be identified with their κ, Z_1 and Z_2; our r_2, σ_2 and τ_2 should be identified with their r, σ and τ; our regularization is defined as $\frac{\lambda}{2n}\|\bullet\|^2$ whereas theirs is defined as $\frac{\lambda}{2p}\|\bullet\|^2$, so to see the equivalence, one needs to make the replacement $\lambda \mapsto \lambda\kappa^{-1}$ above.*

Random permutations and sign flipping. Recall the setup for random permutations and random sign flipping in Section 6. We first verify that the equations (EQs) do apply to these two augmentations in special cases. In view of Theorem 12, the key condition to verify is Assumption 11.

Lemma 54 *Suppose the coordinates of each $Z_1^{(t)}$ are i.i.d. within the group. Then Assumption 11 holds for random permutations.*

Lemma 55 *Suppose* $\mathrm{Var}[Z_1] = \frac{1}{p}I_p$. *Then Assumption 11 holds for random sign flipping.*

Random cropping. Recall the random cropping scheme defined in Section 6. While random cropping does not satisfy Assumption 11, it does satisfy a slightly relaxed notion of Assumption 11:

Lemma 56 *Suppose* $\mathrm{Var}[Z_1] = \frac{1}{p}I_p$. *For random cropping, there exist some $a_1, a_2 > 0$ such that*

$$(i)\ \Sigma_* \;= a_1(\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_1(Z_1)\,,\,Z_1](\Sigma_o^\dagger)^{1/2} \qquad \textit{and} \qquad (ii)\ \Sigma_*^2 \;= a_2\Sigma_*\,. \tag{108}$$

The core CGMT statement — the equivalence of (PO) and (AO) — does hold for random cropping. To see this, notice that Assumption 11 is equivalent to having $a_1 = a_2 = 1$ in (108). We observe that to prove the equivalence of (PO) and (AO) in Lemma 48, Assumption 11 is only critical for showing the independence of the differently projected data matrices, which hold even under the rescaling a_1 and a_2 in (108); see the proof of Lemma 57 below. As such, Meanwhile, a tedious extension of (EQs) also holds for random cropping. Notice that Assumption 11 is used again only in the calculations from (110) onwards in the proof of Lemma 49, which relates (AO) to (SO). There, we only use the idempotency of Σ_* such that Σ_* and $I_p - \Sigma_*$ are projections onto orthogonal subspaces, which simplify many subsequent calculations. If instead (108)(ii) holds, a similar calculation still works by writing $\Sigma_* = \Sigma_1 + \Sigma_2$ and $I_p - \Sigma_* = \Sigma_2' + \Sigma_3$, such that Σ_1, Σ_2 and Σ_3 have mutually orthogonal positive eigenspaces, and Σ_2 and Σ_2' share the same positive eigenspace. This would lead to a system of equations involving $(\sigma_1, \sigma_2, \sigma_3, \tau_1, \tau_2, \tau_3)$ instead of just $(\sigma_1, \sigma_2, \tau_1, \tau_2)$ in (EQs), and we omit the calculations for simplicity.

Appendix N. Proofs for Appendix M

N.1. Proofs for the equivalence of (PO) and (AO)

The next lemma confirms that the projection P_* decouples the different random quantities.

Lemma 57 *Under Assumption 11,* $\mathbf{G}^\Phi P_*^\perp$ *is independent of* $(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi P_*)$.

Proof of Lemma 57 By Gaussianity, to prove independence, it suffices to check that the covariance between the random quantities are zero. We first verify that the covariance between $\mathbf{G}^\Phi P_*^\perp$ and $\mathbf{G}\Sigma_o^{1/2}\beta^*$ is zero, for which it suffices to compute

$$\begin{aligned}\mathrm{Cov}[P_*^\perp G_{11}^\Phi\,,\, G_1^\perp \Sigma_o^{1/2}\beta^*] &= P_*^\perp\, \mathrm{Cov}[(\Sigma^\dagger)^{1/2}\phi_{11}(X_1), (\Sigma_o^\dagger)^{1/2} X_1]\, \Sigma_o^{1/2}\beta^* \\ &= P_*^\perp (\Sigma^\dagger)^{1/2}\, \mathrm{Cov}[\phi_{11}(X_1), X_1]\, (\Sigma_o^\dagger)^{1/2}\Sigma_o^{1/2}\beta^* \\ &= P_*^\perp \Sigma_* \Sigma_o^{1/2}\beta^* \;=\; 0\,.\end{aligned}$$

In the last line, we used Assumption 11(i), and concluded that the covariance evaluates to zero by the definition of P_*. This proves that $\mathbf{G}^\Phi P_*^\perp$ is independent of $\mathbf{G}\Sigma_o^{1/2}\beta^*$.

To check the independence between $\mathbf{G}^\Phi P_*^\perp$ and $\mathbf{G}^\Phi P_*$, we first note that since $\Sigma = \mathrm{Var}[\phi_{11}(X_1)]$, we have

$$\mathrm{Cov}[P_*^\perp G_{11}^\Phi\,,\, P_* G_{11}^\Phi] \;=\; P_*^\perp (\Sigma^\dagger)^{1/2}\, \mathrm{Var}[\phi_{11}(X_1)]\, (\Sigma^\dagger)^{1/2} P_* \;=\; P_*^\perp P_* \;=\; 0\,.$$

We also need to compute

$$\begin{aligned}\mathrm{Cov}[P_*^\perp G_{11}^\Phi\,,\, P_* G_{12}^\Phi] &= P_*^\perp (\Sigma^\dagger)^{1/2}\, \mathrm{Cov}[\phi_{11}(X_1)\,,\, \phi_{12}(X_1)]\, (\Sigma^\dagger)^{1/2} P_* \\ &= P_*^\perp\, \Sigma_*\, P_*\,.\end{aligned}$$

Now note that if $\Sigma_*\Sigma_o^{1/2}\beta^* = 0$, the above evaluates to zero automatically. Otherwise, we have

$$\Sigma_*\, P_* \;=\; \Sigma_*\, \frac{(\Sigma_*\Sigma_o^{1/2}\beta^*)(\Sigma_*\Sigma_o^{1/2}\beta^*)^\intercal}{\|\Sigma_*\Sigma_o^{1/2}\beta^*\|^2} \;=\; P_*\,,$$

where we have used $\Sigma_*^2 = \Sigma_*$ by Assumption 11(ii). This implies

$$\mathrm{Cov}[P_*^\perp G_{11}^\Phi\,,\, P_* G_{12}^\Phi] \;=\; P_*^\perp P_* \;=\; 0\,,$$

which proves that $\mathbf{G}^\Phi P_*^\perp$ is independent of $\mathbf{G}^\Phi P_*$. ■

Lemma 57 suggests that we can apply Theorem 5 to $\mathbf{G}^\Phi P_*^\perp \Sigma^{1/2}$ conditionally on $(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi P_*)$. To facilitate this, the next lemma computes the covariance structure of $\mathbf{G}^\Phi P_*^\perp \Sigma^{1/2}$.

Lemma 58 *For $i, i' \leq m$, $j, j' \leq k$ and $l, l' \leq p$,*

$$\mathrm{Cov}[(\Sigma^{1/2} P_*^\perp G_{ij}^\Phi)_l \,,\, (\Sigma^{1/2} P_*^\perp G_{i'j'}^\Phi)_{l'}] \;=\; (I_{mk})_{ij,i'j'}\,(\Sigma_1)_{l,l'} + (J_{mk})_{ij,i'j'}\,(\Sigma_2)_{l,l'} \;.$$

Moreover, Σ_*, Σ_1 and Σ_2 are all positive semi-definite.

Proof of Lemma 58 For $i, i' \leq m$, $j, j' \leq k$ and $l, l' \leq p$, we have

$$\begin{aligned}
&\mathrm{Cov}[(\Sigma^{1/2} P_*^\perp G_{ij}^\Phi)_l \,,\, (\Sigma^{1/2} P_*^\perp G_{i'j'}^\Phi)_{l'}] \\
&= \mathrm{Cov}[(\Sigma^{1/2} P_*^\perp (\Sigma^\dagger)^{1/2} \phi_{ij}(X_i))_l \,,\, (\Sigma^{1/2} P_*^\perp (\Sigma^\dagger)^{1/2} \phi_{i'j'}(X_{i'}))_{l'}] \\
&= \mathbb{I}\{i = i'\}\mathbb{I}\{j = j'\}\,(\,\Sigma^{1/2} P_*^\perp (\Sigma^\dagger)^{1/2}\,\mathrm{Var}[\phi_{11}(X_1)]\,(\Sigma^\dagger)^{1/2} P_*^\perp \Sigma^{1/2}\,)_{l,l'} \\
&\quad + \mathbb{I}\{i = i'\}\mathbb{I}\{j \neq j'\}\,(\,\Sigma^{1/2} P_*^\perp (\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_{11}(X_1)\,,\,\phi_{12}(X_1)]\,(\Sigma^\dagger)^{1/2} P_*^\perp \Sigma^{1/2}\,)_{l,l'} \\
&\overset{(a)}{=} (I_{mk})_{ij,i'j'}\,(\Sigma^{1/2} P_*^\perp \Sigma^{1/2} - \Sigma^{1/2} P_*^\perp \Sigma_* P_*^\perp \Sigma^{1/2})_{l,l'} + (J_{mk})_{ij,i'j'}\,(\Sigma^{1/2} P_*^\perp \Sigma_* P_*^\perp \Sigma^{1/2})_{l,l'} \\
&= (I_{mk})_{ij,i'j'}\,(\Sigma_1)_{l,l'} + (J_{mk})_{ij,i'j'}\,(\Sigma_2)_{l,l'} \;.
\end{aligned}$$

In (a), we have used that $(\Sigma^\dagger)^{1/2}\,\mathrm{Var}[\phi_{11}(X_1)]\,(\Sigma^\dagger)^{1/2} = I_p$, $(P_*^\perp)^2 = P_*^\perp$ and the definition of Σ_*. This gives the desired formula. Now by the total law of covariance (see e.g. Lemma 41(i) of Huang et al. (2022)),

$$\begin{aligned}
\Sigma_* &= (\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_{11}(X_1)\,,\,P_*^\perp \phi_{12}(X_1)]\,(\Sigma^\dagger)^{1/2} \\
&= (\Sigma^\dagger)^{1/2}\,\mathrm{Var}\,\mathbb{E}[\phi_{11}(X_1)\,|\,X_1]\,(\Sigma^\dagger)^{1/2}
\end{aligned}$$

which is positive semi-definite. This implies that Σ_2 is also positive semi-definite. Moreover, by another total law of variance, we get that

$$\Sigma_* \;\preceq\; (\Sigma^\dagger)^{1/2}\,\mathrm{Var}[\phi_{11}(X_1)]\,(\Sigma^\dagger)^{1/2} \;=\; I_p \;,$$

where $\preceq$ denotes the Loewner partial order on positive semi-definite matrices. This implies that $I_p - \Sigma_*$ is positive semi-definite and so is Σ_1. ■

We are now ready to prove the equivalence of (PO) and (AO).

Proof of Lemma 48 We recall that (PO) can be expressed as

$$\begin{aligned}
\min_{\beta \in S, u \in S_u} \max_{v \in S_v} \; &\frac{1}{mk} \mathbf{1}_{nk}^\top \rho(u) - \frac{1}{mk} \mathbf{y}(\mathbf{G}\Sigma_o^{1/2}\beta^*)^\top u + \frac{\lambda}{2m}\|\beta\|_2^2 + \frac{1}{mk} v^\top u \\
&- \frac{1}{mk} v^\top \mathbf{G}^\Phi P_* \Sigma^{1/2}\beta - \frac{1}{mk} v^\top \mathbf{G}^\Phi P_*^\perp \Sigma^{1/2}\beta \;.
\end{aligned}$$

By Lemma 57, $\mathbf{G}^\Phi P_*^\perp$ is independent of $(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi P_*)$. This allows us to condition on the random variables $(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi P_*)$, apply the CGMT result to $\mathbf{G}^\Phi P_*^\perp \Sigma^{1/2}$, and then marginalize out $(\mathbf{G}\Sigma_o^{1/2}\beta^*, \mathbf{G}^\Phi P_*)$. Notice that the loss is convex-concave in (β, v), the sets of optimization are compact convex, and the variance-covariance structure of $\mathbf{G}^\Phi P_*^\perp \Sigma^{1/2}$ is given by Lemma 58, which satisfies the condition of Theorem 5 with $M = 2$. The conclusions of Theorem 5 therefore hold for (PO) and (AO). ■

N.2. Proof of Lemma 49: Equivalence between (PO), (AO) and (SO)

The calculations are mostly similar to that of Salehi et al. (2019), so we focus on highlighting the differences in the proof.

Analyzing the auxiliary optimization. We first notice that, other than the regularization term $\|\beta\|_2^2$, β appears in the loss only through $\Sigma^{1/2}\beta$, $\Sigma_1^{1/2}\beta$ and $\Sigma_2^{1/2}\beta$, where

$$\Sigma_1 \; = \Sigma^{1/2} P_*^\perp (I_p - \Sigma_*) P_*^\perp \Sigma^{1/2} \qquad\qquad \text{and} \qquad\qquad \Sigma_2 \; = \Sigma^{1/2} P_*^\perp \Sigma^{1/2} \; .$$

Therefore it suffices to restrict the set of minimization, $\beta \in S$, to the intersection of S and the positive eigenspace of Σ. Define the projection to the positive eigenspace of Σ as $P_\Sigma := \Sigma^\dagger \Sigma$, which allows us to rewrite the auxiliary optimization as

$$\begin{aligned}\min_{\beta \in S, u \in S_u} \max_{v \in S_v} & \frac{1}{mk} \mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk} \mathbf{y}^\top u + \frac{\lambda}{2m} \|P_\Sigma \beta\|_2^2 + \frac{1}{mk} v^\top (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta) \\ & - \frac{1}{mk} v^\top \mathbf{h}_1 \|\beta\|_{\Sigma_1} + \frac{1}{mk} \|v\| \mathbf{g}_1^\top \Sigma_1^{1/2} \beta \\ & - \frac{1}{mk^{3/2}} v^\top J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2} + \frac{1}{mk} \|v\|_{J_{mk}} \mathbf{g}_2^\top \Sigma_2^{1/2} \beta \; . \end{aligned} \tag{109}$$

For simplicity, we have abbreviated $\mathbf{y} \equiv \mathbf{y}(\mathbf{G} \Sigma_o^{1/2} \beta^*)$.

Salehi et al. (2019) showed that, under their CGMT result (analogous to our Theorem 5(i) and Theorem 5(ii)), the minimum and maximum can be exchanged in the auxiliary optimization in an asymptotic sense since they can be exchanged in the primary optimization. Throughout the analysis of AO, we will highlight explicitly where such flipping is done, and in the case where the min-max theorem is not applicable, we defer a rigorous justification to the end of the proof.

For simplicity of notation, given a matrix $A \in \mathbb{R}^{d' \times d}$ and a subset $S \in \mathbb{R}^d$, we also write $A(S) = \{Av \,|\, v \in S\}$ for short.

Maximizing over $v \in S_v \subset \mathbb{R}^{mk}$. Consider the projection matrix $P_{mk} := \frac{1}{k} J_{mk}$ and write $P_{mk}^\perp = I_{mk} - P_{mk}$. Notice also that $\|\bullet\|_{J_{mk}} = \sqrt{k} \|P_{mk}(\bullet)\|$. Then the maximization over v can be re-expressed as

$$\begin{aligned}\max_{P_{mk}^\perp v \in P_{mk}^\perp (S_v)} \max_{P_{mk} v \in P_{mk}(S_v)} \; & \frac{1}{mk} v^\top P_{mk} (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2}) \\ & + \frac{1}{mk} v^\top P_{mk}^\perp (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2}) \\ & + \frac{1}{mk} \|P_{mk} v\| \mathbf{g}_1 \Sigma_1^{1/2} \beta + \frac{1}{mk} \|P_{mk}^\perp v\| \mathbf{g}_1^\top \Sigma_1^{1/2} \beta + \frac{1}{m\sqrt{k}} \|P_{mk} v\| \mathbf{g}_2^\top \Sigma_2^{1/2} \beta \; . \end{aligned}$$

Maximizing the above over $P_{mk} v$ and $P_{mk}^\perp v$ separately, choosing each vector to be what it multiplies and writing $r_1 = \|P_{mk}^\perp v\| / \sqrt{mk}$ and $r_2 = \|P_{mk} v\| / \sqrt{mk}$ (analogous to (44) – (45) in Salehi et al. (2019)), the above can be rewritten as

$$\begin{aligned}\max_{(r_1, r_2) \in S_{r_1} \times S_{r_2}} \; & r_1 \Big(\frac{1}{\sqrt{mk}} \mathbf{g}_1^\top \Sigma_1^{1/2} \beta + \frac{1}{\sqrt{mk}} \Big\| P_{mk}^\perp (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2}) \Big\| \Big) \\ & + r_2 \Big(\frac{1}{\sqrt{mk}} \mathbf{g}_1^\top \Sigma_1^{1/2} \beta + \frac{1}{\sqrt{m}} \mathbf{g}_2^\top \Sigma_2^{1/2} \beta \\ & \qquad + \frac{1}{\sqrt{mk}} \Big\| P_{mk} (u - \mathbf{G}^\Phi P_* \Sigma^{1/2} \beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk} \mathbf{h}_2 \|\beta\|_{\Sigma_2}) \Big\| \Big) \; , \end{aligned}$$

where we have denoted

$$S_{r_1} := \Big\{ \frac{1}{\sqrt{mk}} \|P_{mk}^{\perp} v\| \,\Big|\, v \in S_v \Big\} \qquad \text{and} \qquad S_{r_2} := \Big\{ \frac{1}{\sqrt{mk}} \|P_{mk} v\| \,\Big|\, v \in S_v \Big\} \,.$$

Substituting this into (109) yields

$$\begin{aligned} \min_{\substack{\beta\in S\\ u\in S_u}} \max_{(r_1,r_2)\in S_{r_1}\times S_{r_2}} \; & \frac{1}{mk}\mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk}\mathbf{y}^\top u + \frac{\lambda}{2m}\|P_\Sigma \beta\|_2^2 + \frac{r_1+r_2}{\sqrt{mk}}\mathbf{g}_1^\top \Sigma_1^{1/2}\beta + \frac{r_2}{\sqrt{n}}\mathbf{g}_2^\top \Sigma_2^{1/2}\beta \\ & + \frac{r_1}{\sqrt{mk}}\Big\|P_{mk}^{\perp}\big(u - \mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk}\mathbf{h}_2 \|\beta\|_{\Sigma_2}\big)\Big\| \\ & + \frac{r_2}{\sqrt{mk}}\Big\|P_{mk}\big(u - \mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk}\mathbf{h}_2 \|\beta\|_{\Sigma_2}\big)\Big\| \,. \end{aligned}$$

Minimizing over $\beta \in S$. As with (47) of Salehi et al. (2019), we introduce new variables $\mu, w \in \mathbb{R}^p$ to replace β in the regularization term via the Lagrange multiplier method applied to the constraint $P_\Sigma \mu = P_\Sigma \beta$:

$$\min_{\substack{\beta\in S\\ u\in S_u\\ \mu\in S}} \max_{\substack{w\in\mathbb{R}^p\\ (r_1,r_2)\in S_{r_1}\times S_{r_2}}} \mathcal{L}_1(\beta, u, \mu, w, r_1, r_2) \,,$$

where

$$\begin{aligned} \mathcal{L}_1(\beta, u, \mu, w, r_1, r_2) \;:=\; & \frac{1}{mk}\mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk}\mathbf{y}^\top u + \frac{\lambda}{2m}\|P_\Sigma \mu\|_2^2 + \frac{1}{p} w^\top P_\Sigma(\mu - \beta) \\ & + \frac{r_1+r_2}{\sqrt{mk}}\mathbf{g}_1^\top \Sigma_1^{1/2}\beta + \frac{r_2}{\sqrt{m}}\mathbf{g}_2^\top \Sigma_2^{1/2}\beta \\ & + \frac{r_1}{\sqrt{mk}}\Big\|P_{mk}^{\perp}\big(u - \mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk}\mathbf{h}_2 \|\beta\|_{\Sigma_2}\big)\Big\| \\ & + \frac{r_2}{\sqrt{mk}}\Big\|P_{mk}\big(u - \mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta - \mathbf{h}_1 \|\beta\|_{\Sigma_1} - \frac{1}{\sqrt{k}} J_{mk}\mathbf{h}_2 \|\beta\|_{\Sigma_2}\big)\Big\| \,. \end{aligned} \tag{110}$$

To minimize over $\beta \in S$, we first swap the order of $\min_{\beta\in S}$ and $\max_{w\in\mathbb{R}^p,\,(r_1,r_2)\in S_{r_1}\times S_{r_2}}$. Notice that the β-dependence in the loss comes from $P_\Sigma\beta$, $\Sigma_1^{1/2}\beta$, $\Sigma_2^{1/2}\beta$ and $\mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta$. Writing $\tilde\beta = \sqrt{p}\, P_* \Sigma^{1/2}\beta$, $\tilde\beta^\perp = \sqrt{p}\, P_*^\perp \Sigma^{1/2}\beta$, $v(\beta^*) := \sqrt{p}\, \Sigma_* \Sigma_o^{1/2}\beta^*$ and $\kappa_* := \|\Sigma_*\Sigma_o^{1/2}\beta^*\|$, we can express

$$\begin{aligned} \Sigma_1^{1/2}\beta &= (\Sigma^{1/2} P_*^\perp (I_p - \Sigma_*) P_*^\perp \Sigma^{1/2})^{1/2}\beta \overset{(a)}{=} \frac{1}{\sqrt{p}}(I_p - \Sigma_*)\tilde\beta^\perp \,, \\ \Sigma_2^{1/2}\beta &= (\Sigma^{1/2} P_*^\perp \Sigma_* P_*^\perp \Sigma^{1/2})^{1/2}\beta \overset{(b)}{=} \frac{1}{\sqrt{p}}\Sigma_*\tilde\beta^\perp \,, \\ \mathbf{G}^{\Phi} P_* \Sigma^{1/2}\beta &= \underbrace{\frac{1}{\sqrt{p}}\mathbf{G}^{\Phi} v(\beta^*)}_{=:\kappa_*\mathbf{q}} \times \underbrace{\frac{v(\beta^*)^\top \Sigma^{1/2}\beta}{\sqrt{p}\,\kappa_*^2}}_{=:\alpha(\tilde\beta)} \,, \\ P_\Sigma\beta &= \frac{1}{\sqrt{p}}(\Sigma^\dagger)^{1/2}\tilde\beta + \frac{1}{\sqrt{p}}(\Sigma^\dagger)^{1/2}\tilde\beta^\perp = \frac{\alpha(\tilde\beta)}{\sqrt{p}}(\Sigma^\dagger)^{1/2} v(\beta^*) + \frac{1}{\sqrt{p}}(\Sigma^\dagger)^{1/2}\tilde\beta^\perp \,. \end{aligned}$$

In (a) and (b) above, we have used Assumption 11(ii) to note that $I_p - \Sigma_*$ and Σ_* are both idempotent. This allows us to express all β-dependent terms in terms of $\alpha(\tilde\beta)$ and $\tilde\beta^\perp$, where $\tilde\beta$ and $\tilde\beta^\perp$ are orthogonal and can be optimized separately. Therefore, the optimization can be rewritten as

$$\begin{aligned}\min_{\substack{u\in S_u\\ \mu\in S\\ \alpha\in S^{\alpha}}} \; \max_{\substack{w\in\mathbb{R}^p\\ (r_1,r_2)\in S_{r_1}\times S_{r_2}}} \; \min_{\tilde{\beta}^{\perp}\in\tilde{S}^{\perp}} \; & \frac{1}{mk}\mathbf{1}_{mk}^{\top}\rho(u) - \frac{1}{mk}\mathbf{y}^{\top}u + \frac{\lambda}{2m}\|P_{\Sigma}\mu\|_2^2 + \frac{1}{p}w^{\top}P_{\Sigma}\mu \\ & - \frac{\alpha}{p\sqrt{p}}\, w^{\top}(\Sigma^{\dagger})^{1/2}v(\beta^*) - \frac{1}{p\sqrt{p}}w^{\top}(\Sigma^{\dagger})^{1/2}\tilde{\beta}^{\perp} \\ & + \frac{r_1+r_2}{\sqrt{mkp}}\mathbf{g}_1^{\top}(I_p-\Sigma_*)\tilde{\beta}^{\perp} + \frac{r_2}{\sqrt{mp}}\mathbf{g}_2^{\top}\Sigma_*\tilde{\beta}^{\perp} \\ & + \frac{r_1}{\sqrt{mk}}\Big\|P_{mk}^{\perp}\Big(u-\kappa_*\alpha\mathbf{q}-\frac{1}{\sqrt{p}}\mathbf{h}_1\|(I_p-\Sigma_*)\tilde{\beta}^{\perp}\| - \frac{1}{\sqrt{pk}}J_{mk}\mathbf{h}_2\|\Sigma_*\tilde{\beta}^{\perp}\|\Big)\Big\| \\ & + \frac{r_2}{\sqrt{mk}}\Big\|P_{mk}\Big(u-\kappa_*\alpha\mathbf{q}-\frac{1}{\sqrt{p}}\mathbf{h}_1\|(I_p-\Sigma_*)\tilde{\beta}^{\perp}\| - \frac{1}{\sqrt{pk}}J_{mk}\mathbf{h}_2\|\Sigma_*\tilde{\beta}^{\perp}\|\Big)\Big\|\,, \end{aligned} \tag{111}$$

where we have defined the sets

$$S^{\alpha} := \{\tfrac{v(\beta^*)^{\top}\Sigma^{1/2}\beta}{\sqrt{p}\,\kappa_*^2}\,|\beta\in S\} \qquad \text{and} \qquad \tilde{S}^{\perp} := \{\sqrt{p}\,P_*^{\perp}\Sigma^{1/2}\beta\,|\,\beta\in S\}\,.$$

Note that we have moved the minimization over α to the outmost part of the loss. The steps so far are analogous to (46) – (47) of Salehi et al. (2019). Before proceeding, we notice that since $I_p-\Sigma_*$ and Σ_* are symmetric and idempotent by Assumption 11(ii), they are in fact projection matrices onto two orthogonal subspaces. Therefore to optimize the above over $\tilde{\beta}^{\perp}$, it suffices to do so over $(I_p-\Sigma_*)\tilde{\beta}^{\perp}$ and $\Sigma_*\tilde{\beta}^{\perp}$ individually. Moreover, when optimizing over each of the projected $\tilde{\beta}^{\perp}$'s, the optimization takes exactly the same form as (47) of Salehi et al. (2019). Similar to them, we introduce

$$\sigma_1 := \frac{1}{\sqrt{p}}\|(I_p-\Sigma_*)\tilde{\beta}^{\perp}\| \;\in\; S_{\sigma_1} \qquad \text{and} \qquad \sigma_2 := \frac{1}{\sqrt{p}}\|\Sigma_*\tilde{\beta}^{\perp}\| \;\in\; S_{\sigma_2}\,,$$

where $S_{\sigma_1} := \{\|(I_p-\Sigma_*)P_*^{\perp}\Sigma^{1/2}\beta\|\,|\beta\in S\}$ and $S_{\sigma_2} := \{\|\Sigma_*P_*^{\perp}\Sigma^{1/2}\beta\|\,|\beta\in S\}$ are both subsets of non-negative real numbers, as well as the auxiliary variables $\nu_1,\nu_2,\tau_1,\tau_2\geq 0$. We also take note of the fact that

$$\begin{aligned}(I_p-\Sigma_*)\tilde{\beta}^{\perp} &\in S_{\Sigma}^{\perp} := \{\sqrt{p}\,(I_p-\Sigma_*)P_*^{\perp}\Sigma^{1/2}\beta\,|\,\beta\in S\}\,, \\ \Sigma_*\tilde{\beta}^{\perp} &\in S_{\Sigma} := \{\sqrt{p}\,\Sigma_*P_*^{\perp}P_{\Sigma}\Sigma^{1/2}\beta\,|\,\beta\in S\}\,,\end{aligned}$$

and denote the projection onto $\mathrm{span}(S^{\perp}_{\Sigma})$ as $P_{S^{\perp}_{\Sigma}}$ and the projection onto $\mathrm{span}(S_{\Sigma})$ as $P_{S_{\Sigma}}$. Then by the same algebra from (47) – (49) of Salehi et al. (2019), we obtain

$$\begin{aligned}\min_{\substack{u\in S_u\\ \mu\in S\\ \alpha\in S^{\alpha}\\ (\sigma_1,\sigma_2)\in S_{\sigma_1}\times S_{\sigma_2}\\ \nu_1,\nu_2\geq 0}} \quad \max_{\substack{w\in\mathbb{R}^p\\ (r_1,r_2)\in S_{r_1}\times S_{r_2}\\ \tau_1,\tau_2\geq 0}} \quad & \frac{1}{mk}\mathbf{1}_{mk}^{\top}\rho(u) - \frac{1}{mk}\mathbf{y}^{\top}u + \frac{\lambda}{2m}\|P_{\Sigma}\mu\|_2^2 + \frac{1}{p}w^{\top}P_{\Sigma}\mu \\ & - \frac{\alpha}{p\sqrt{p}}\, w^{\top}(\Sigma^{\dagger})^{1/2}v(\beta^*) \\ & - \frac{\sigma_1}{2\tau_1} - \frac{\sigma_1\tau_1}{2}\left\|P_{S^{\perp}_{\Sigma}}\Big(\frac{r_1+r_2}{\sqrt{mk}}\mathbf{g}_1 - \frac{1}{p}(I_p-\Sigma_*)(\Sigma^{\dagger})^{1/2}w\Big)\right\|^2 \\ & - \frac{\sigma_2}{2\tau_2} - \frac{\sigma_2\tau_2}{2}\left\|P_{S_{\Sigma}}\Big(\frac{r_2}{\sqrt{n}}\mathbf{g}_2 - \frac{1}{p}\Sigma_*(\Sigma^{\dagger})^{1/2}w\Big)\right\|^2 \\ & + \frac{r_1}{2\nu_1} + \frac{r_1\nu_1}{2mk}\left\|P^{\perp}_{mk}\Big(u - \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}}J_{mk}\mathbf{h}_2\Big)\right\|^2 \\ & + \frac{r_2}{2\nu_2} + \frac{r_2\nu_2}{2mk}\left\|P_{mk}\Big(u - \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}}J_{mk}\mathbf{h}_2\Big)\right\|^2 . \end{aligned} \tag{112}$$

Note that we have moved the maximization over w, r_1, r_2 to be inside the minimization over ν_1, ν_2, σ_1 and σ_2. Note also that $P^{\perp}_{mk}\sigma_2 J_{mk}\mathbf{h}_2$ evaluates to zero, but we keep this term for the ease of computation later. We also remark that ν_1 can be restricted to be in a compact set $\{\big\|P^{\perp}_{mk}(u - \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}}J_{mk}\mathbf{h}_2)\big\|\,\big|\,u\in S_u\}$ for the purpose of flipping minimization and maximization, and so are ν_2, τ_1, τ_2, but we do not do so for notational simplicity.

Maximization over $w\in\mathbb{R}^p$. We first derive some useful relationships between the different projection matrices introduced so far: By the definition of Σ_*, we have

$$\Sigma_* P_{\Sigma} \;=\; (\Sigma^{\dagger})^{1/2}\mathrm{Cov}[\phi_{11}(X_1), \phi_{11}(X_2)](\Sigma^{\dagger})^{1/2}P_{\Sigma} \;=\; \Sigma_* \,. \tag{113}$$

Also by the definition of P_* and the idempotency of Σ_*,

$$P_*\Sigma_* \;=\; \Sigma_* P_* \;=\; \begin{cases}\Sigma_*\,\dfrac{(\Sigma_*\Sigma_o^{1/2}\beta^*)(\Sigma_*\Sigma_o^{1/2}\beta^*)^{\top}}{\|\Sigma_*\Sigma_o^{1/2}\beta^*\|^2} \;=\; P_* & \text{if } \Sigma_*\Sigma_o^{1/2}\beta^* \neq 0 \\ \Sigma_*\times 0 \;=\; P_* & \text{otherwise}\;\;. \end{cases} \tag{114}$$

This implies that

$$\begin{aligned} S^{\perp}_{\Sigma} \;&= \{\sqrt{p}\,(I_p-\Sigma_*)(I_p-P_*)\Sigma^{1/2}\beta\,|\,\beta\in S\} \;=\; \{\sqrt{p}\,(I_p-\Sigma_*)\Sigma^{1/2}\beta\,|\,\beta\in S\}\,, \\ S_{\Sigma} \;&= \{\Sigma_*(I_p-P_*)\Sigma^{1/2}\beta\,|\,\beta\in S\} \;=\; \{(\Sigma_*-P_*)\Sigma^{1/2}\beta\,|\,\beta\in S\}\,, \end{aligned}$$

and combining these with the assumption that $\mathrm{span}(S) = \mathbb{R}^p$, we can express

$$P_{S^{\perp}_{\Sigma}} = I_p - \Sigma_* \qquad\qquad \text{and} \qquad\qquad P_{S_{\Sigma}} = \Sigma_* - P_* \,. \tag{115}$$

This in turn implies that

$$P_{S^{\perp}_{\Sigma}}(I_p-\Sigma_*) = P_{S^{\perp}_{\Sigma}}\,, \qquad P_{S_{\Sigma}}\Sigma_* = P_{S^{\perp}_{\Sigma}}\,, \qquad P_{S^{\perp}_{\Sigma}}P_{S_{\Sigma}} = P_*P_{S^{\perp}_{\Sigma}} = P_*P_{S_{\Sigma}} = 0\,, \tag{116}$$

and that

$$\begin{aligned}
P_\Sigma w &= \Sigma^{1/2}(\Sigma^\dagger)^{1/2} w \\
&= \Sigma^{1/2} P_*(\Sigma^\dagger)^{1/2} w + \Sigma^{1/2} P_{S_\Sigma^\perp}(\Sigma^\dagger)^{1/2} w + \Sigma^{1/2} P_{S_\Sigma}(\Sigma^\dagger)^{1/2} w \\
&= \Sigma^{1/2} \frac{v(\beta^*)v(\beta^*)^\intercal}{p\kappa_*^2}(\Sigma^\dagger)^{1/2} w + \Sigma^{1/2} P_{S_\Sigma^\perp}(\Sigma^\dagger)^{1/2} w + \Sigma^{1/2} P_{S_\Sigma}(\Sigma^\dagger)^{1/2} w \;.
\end{aligned}$$

Substituting these into (112), we obtain

$$\begin{aligned}
\min_{\substack{u \in S_u \\ \mu \in S \\ \alpha \in S^\alpha \\ (\sigma_1,\sigma_2) \in S_{\sigma_1} \times S_{\sigma_2} \\ \nu_1,\nu_2 \geq 0}} \quad \max_{\substack{w \in \mathbb{R}^p \\ (r_1,r_2) \in S_{r_1} \times S_{r_2} \\ \tau_1,\tau_2 \geq 0}} \quad & \frac{1}{mk} \mathbf{1}_{mk}^\intercal \rho(u) - \frac{1}{mk} \mathbf{y}^\intercal u + \frac{\lambda}{2n} \|P_\Sigma \mu\|_2^2 \\
& + \Big(\frac{1}{p^2 \kappa_*^2} \mu^\intercal \Sigma^{1/2} v(\beta^*) - \frac{\alpha}{p\sqrt{p}} \Big) v(\beta^*)^\intercal P_* (\Sigma^\dagger)^{1/2} w \\
& + \frac{1}{p} (\Sigma^{1/2}\mu)^\intercal P_{S_\Sigma^\perp} (\Sigma^\dagger)^{1/2} w + \frac{1}{p} (\Sigma^{1/2}\mu)^\intercal P_{S_\Sigma} (\Sigma^\dagger)^{1/2} w \\
& - \frac{\sigma_1}{2\tau_1} - \frac{\sigma_1 \tau_1}{2} \Big\| P_{S_\Sigma^\perp} \Big(\frac{r_1 + r_2}{\sqrt{mk}} \mathbf{g}_1 - \frac{1}{p} (\Sigma^\dagger)^{1/2} w \Big) \Big\|^2 \\
& - \frac{\sigma_2}{2\tau_2} - \frac{\sigma_2 \tau_2}{2} \Big\| P_{S_\Sigma} \Big(\frac{r_2}{\sqrt{n}} \mathbf{g}_2 - \frac{1}{p} (\Sigma^\dagger)^{1/2} w \Big) \Big\|^2 \\
& + \frac{r_1}{2\nu_1} + \frac{r_1 \nu_1}{2mk} \Big\| P_{mk}^\perp \big(u - \kappa_* \alpha \mathbf{q} - \sigma_1 \mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}} J_{mk} \mathbf{h}_2 \big) \Big\|^2 \\
& + \frac{r_2}{2\nu_2} + \frac{r_2 \nu_2}{2mk} \Big\| P_{mk} \big(u - \kappa_* \alpha \mathbf{q} - \sigma_1 \mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}} J_{mk} \mathbf{h}_2 \big) \Big\|^2 \;. \qquad (117)
\end{aligned}$$

To optimize the above over w, it again suffices to optimize over three mutually orthogonal vectors $P_*(\Sigma^\dagger)^{1/2} w$, $P_{S_\Sigma^\perp}(\Sigma^\dagger)^{1/2} w$ and $P_{S_\Sigma}(\Sigma^\dagger)^{1/2} w$. The optimization over $P_*(\Sigma^\dagger)^{1/2} w$ is exactly analogous to the optimization over $\mathbf{Pw}$ in (49) of Salehi et al. (2019), whereas the optimization over the other two vectors are exactly analogous to that over $\mathbf{P}^\perp \mathbf{w}$ in (49) of Salehi et al. (2019). Therefore by the exact same completion-of-squares argument as in (49) – (51) in Salehi et al. (2019) but without taking the asymptotic approximation, the optimization becomes

$$\begin{aligned}
\min_{\substack{u \in S_u \\ \mu \in S \\ \alpha \in S^\alpha \\ (\sigma_1,\sigma_2) \in S_{\sigma_1} \times S_{\sigma_2} \\ \nu_1,\nu_2 \geq 0 \\ \frac{1}{\sqrt{p}} \mu^\intercal \Sigma^{1/2} v(\beta^*) = \alpha \kappa_*^2}} \quad \max_{\substack{(r_1,r_2) \in S_{r_1} \times S_{r_2} \\ \tau_1,\tau_2 \geq 0}} \quad & \frac{1}{mk} \mathbf{1}_{mk}^\intercal \rho(u) - \frac{1}{mk} \mathbf{y}^\intercal u + \frac{\lambda}{2m} \|P_\Sigma \mu\|_2^2 - \frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} \\
& + \frac{1}{2\sigma_1 \tau_1} \|P_{S_\Sigma^\perp} \Sigma^{1/2} \mu\|^2 + \frac{r_1 + r_2}{\sqrt{mk}} \mathbf{g}_1^\intercal P_{S_\Sigma^\perp} \Sigma^{1/2} \mu \\
& + \frac{1}{2\sigma_2 \tau_2} \|P_{S_\Sigma} \Sigma^{1/2} \mu\|^2 + \frac{r_2}{\sqrt{m}} \mathbf{g}_2^\intercal P_{S_\Sigma} \Sigma^{1/2} \mu \\
& + \frac{r_1 \nu_1}{2mk} \Big\| P_{mk}^\perp \big(u - \kappa_* \alpha \mathbf{q} - \sigma_1 \mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}} J_{mk} \mathbf{h}_2 \big) \Big\|^2 \\
& + \frac{r_2 \nu_2}{2mk} \Big\| P_{mk} \big(u - \kappa_* \alpha \mathbf{q} - \sigma_1 \mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}} J_{mk} \mathbf{h}_2 \big) \Big\|^2 \;, \qquad (118)
\end{aligned}$$

Rewriting the minimization over $\mu \in S$. We now flip the order of optimization such that we can perform the minimization over μ first. This involves computing

$$\begin{aligned} \min_{\mu \in S} \ & \frac{\lambda}{2n}\|P_\Sigma \mu\|_2^2 + \frac{1}{2\sigma_1\tau_1}\|P_{S_\Sigma^\perp}\Sigma^{1/2}\mu\|^2 + \frac{r_1+r_2}{\sqrt{mk}}\mathbf{g}_1^\intercal P_{S_\Sigma^\perp}\Sigma^{1/2}\mu \\ & + \frac{1}{2\sigma_2\tau_2}\|P_{S_\Sigma}\Sigma^{1/2}\mu\|^2 + \frac{r_2}{\sqrt{n}}\mathbf{g}_2^\intercal P_{S_\Sigma}\Sigma^{1/2}\mu \qquad \text{s.t. } \frac{1}{\sqrt{p}}\mu^\intercal\Sigma^{1/2}v(\beta^*) = \alpha\kappa_*^2 \ . \end{aligned} \tag{119}$$

Recall from (115) that $P_{S_\Sigma^\perp} = I_p - \Sigma_*$ and $P_{S_\Sigma} = \Sigma_* - P_*$. Denote

$$\begin{aligned} \tilde{\Sigma}^c_{\sigma,\tau} &:= \frac{1}{2\sigma_1\tau_1}(P_\Sigma - \Sigma_*) + \frac{1}{2\sigma_2\tau_2}(\Sigma_* - P_*) \ , \\ \tilde{\mathbf{g}}^c &:= -\frac{r_1+r_2}{\sqrt{mk}}(P_\Sigma - \Sigma_*)\mathbf{g}_1 - \frac{r_2}{\sqrt{m}}(\Sigma_* - P_*)\mathbf{g}_2 \ . \end{aligned}$$

Then the problem comes

$$\min_{\mu \in S} \frac{\lambda}{2n}\|P_\Sigma\mu\|_2^2 + (\Sigma^{1/2}\mu)^\intercal \tilde{\Sigma}^c_{\sigma,\tau}(\Sigma^{1/2}\mu) - (\tilde{\mathbf{g}}^c)^\intercal\Sigma^{1/2}\mu \qquad \text{s.t. } \frac{1}{\sqrt{p}}\mu^\intercal\Sigma^{1/2}v(\beta^*) = \alpha\kappa_*^2 \ .$$

By (114) and (113), $P_\Sigma P_* = P_\Sigma\Sigma_* P_* = \Sigma_* P_* = P_*$ and by (115), $P_{S_\Sigma^\perp} = I_p - \Sigma_*$ and $P_{S_\Sigma} = \Sigma_* - P_*$. Then by a similar argument as (117), we may express

$$P_\Sigma \ = \ P_\Sigma(P_* + P_{S_\Sigma^\perp} + P_{S_\Sigma}) \ = \ P_* + (P_\Sigma - \Sigma_*) + (\Sigma_* - P_*) \ , \tag{120}$$

where P_*, $P_\Sigma - \Sigma_*$ and $\Sigma_* - P_*$ are projections onto mutually orthogonal subspaces. Meanwhile, recalling the definition of $\tilde{\Sigma}_{\sigma,\tau}$ and $\tilde{\mathbf{g}}$, we can express

$$\begin{aligned} \tilde{\Sigma}_{\sigma,\tau} &= \frac{1}{2\sigma_1\tau_1}(P_\Sigma - \Sigma_*) + \frac{1}{2\sigma_2\tau_2}\Sigma_* \ = \ \tilde{\Sigma}^c_{\sigma,\tau} + \frac{1}{2\sigma_2\tau_2}P_* \ , \\ \tilde{\mathbf{g}} &= -\frac{r_1+r_2}{\sqrt{mk}}(P_\Sigma - \Sigma_*)\mathbf{g}_1 - \frac{r_2}{\sqrt{m}}\Sigma_*\mathbf{g}_2 \ = \ \tilde{\mathbf{g}}^c - \frac{r_2}{\sqrt{m}}P_*\mathbf{g}_2 \ . \end{aligned}$$

Recalling also that $P_* = v(\beta^*)v(\beta^*)^\intercal/(p\kappa_*^2)$, we can write

$$\begin{aligned} & (\Sigma^{1/2}\mu)^\intercal\tilde{\Sigma}^c_{\sigma,\tau}(\Sigma^{1/2}\mu) - (\tilde{\mathbf{g}}^c)^\intercal\Sigma^{1/2}\mu \\ &= (\Sigma^{1/2}\mu)^\intercal\tilde{\Sigma}_{\sigma,\tau}(\Sigma^{1/2}\mu) - \tilde{\mathbf{g}}^\intercal\Sigma^{1/2}\mu - \frac{1}{2\sigma_2\tau_2}(\Sigma^{1/2}\mu)^\intercal P_*(\Sigma^{1/2}\mu) + \frac{r_2}{\sqrt{n}}(P_*\mathbf{g}_2)^\intercal\Sigma^{1/2}\mu \\ &= (\Sigma^{1/2}\mu)^\intercal\tilde{\Sigma}_{\sigma,\tau}(\Sigma^{1/2}\mu) - \tilde{\mathbf{g}}^\intercal\Sigma^{1/2}\mu - \frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2} + \frac{r_2}{\sqrt{m}}\mathbf{g}_2^\intercal P_*\Sigma^{1/2}\mu \ . \end{aligned}$$

Now using a Lagrange multiplier θ to remove the constraint, the optimization becomes

$$\begin{aligned} \min_{\mu\in S}\max_{\theta\in\mathbb{R}} \ & \frac{\lambda}{2n}\|P_\Sigma\mu\|_2^2 + (\Sigma^{1/2}\mu)^\intercal\tilde{\Sigma}_{\sigma,\tau}(\Sigma^{1/2}\mu) - \Big(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)\Big)^\intercal\Sigma^{1/2}\mu \\ & - \frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2} + \frac{r_2}{\sqrt{m}}\mathbf{g}_2^\intercal P_*\Sigma^{1/2}\mu + \alpha\theta\kappa_*^2 \ . \end{aligned}$$

Since the problem is convex-concave, we can apply the min-max theorem of Rockafellar (1970) to flip the order of minimum and maximum. Doing this together with a completion of squares, we obtain

$$\max_{\theta\in\mathbb{R}} M_{\mathbf{g},\sigma,\tau,\theta} - \frac{1}{4}\Big\|(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}\big(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)\big)\Big\|^2 - \frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2} + \alpha\theta\kappa_*^2 \ , \tag{121}$$

where we have denoted the Moreau envelope like term

$$\begin{aligned} M_{\mathbf{g},\sigma,\tau,\theta} \;:=\; \min_{\mu\in S}\; & \frac{\lambda}{2n}\|P_\Sigma \mu\|_2^2 + \Big\|\tilde{\Sigma}_{\sigma,\tau}^{1/2}(\Sigma^{1/2}\mu) - \frac{1}{2}(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}\big(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)\big)\Big\|^2 \\ & + \frac{r_2}{\sqrt{m}}\mathbf{g}_2^\top P_* \Sigma^{1/2}\mu \;. \end{aligned}$$

Rewriting the minimization over $u \in S_u$. Meanwhile, the minimization over $u \in S_u$ involves

$$\min_{u\in S_u} \frac{1}{mk}\mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk}\mathbf{y}^\top u + \frac{r_1\nu_1}{2mk}\Big\|P_{mk}^\perp(u - \tilde{\mathbf{h}}_{\alpha,\sigma})\Big\|^2 + \frac{r_2\nu_2}{2mk}\Big\|P_{mk}(u - \tilde{\mathbf{h}}_{\alpha,\sigma})\Big\|^2 \;, \tag{122}$$

where we have used the shorthand $\tilde{\mathbf{h}}_{\alpha,\sigma} = \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}}J_{mk}\mathbf{h}_2$. Recall that by definition, $\mathbf{y} = P_{mk}\mathbf{y}$ since $\mathbf{y}$ is a length-mk vector formed by k-fold repetitions of m entries. Then we can re-express the loss above as

$$\begin{aligned} & \min_{u\in S_u} \frac{1}{mk}\mathbf{1}_{mk}^\top \rho(u) - \frac{1}{mk}\mathbf{y}^\top P_{mk}u + \frac{r_1\nu_1}{2mk}\Big\|P_{mk}^\perp(u - \tilde{\mathbf{h}}_{\alpha,\sigma})\Big\|^2 + \frac{r_2\nu_2}{2mk}\Big\|P_{mk}(u - \tilde{\mathbf{h}}_{\alpha,\sigma})\Big\|^2 \\ = & \min_{u\in S_u} \frac{1}{mk}\mathbf{1}_{mk}^\top \rho(P_{mk}u + P_{mk}^\perp u) + \frac{r_2\nu_2}{2mk}\Big\|P_{mk}\big(u - \frac{1}{r_2\nu_2}\mathbf{y} - \tilde{\mathbf{h}}_{\alpha,\sigma}\big)\Big\|^2 + \frac{r_1\nu_1}{2mk}\Big\|P_{mk}^\perp(u - \tilde{\mathbf{h}}_{\alpha,\sigma})\Big\|^2 \\ & \qquad - \frac{1}{2r_2\nu_2 mk}\Big\|P_{mk}\mathbf{y}\Big\|^2 - \frac{1}{mk}\mathbf{y}^\top P_{mk}\tilde{\mathbf{h}}_{\alpha,\sigma} \;. \end{aligned}$$

The loss can therefore be minimized separately in $P_{mk}u$ and $P_{mk}^\perp u$. Recall that for a function $f : \mathcal{S} \to \mathbb{R}$ and $\mathcal{S} \subseteq \mathbb{R}^{mk}$, we defined the Moreau envelope

$$\mathcal{M}_S(f; v, t) \;:=\; \min_{x\in\mathcal{S}} f(x) + \frac{1}{2t}\|x - v\|_2^2 \;,$$

Also recall the definition

$$\begin{aligned} M^\perp_{\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu}(\tilde{u}) \;&:=\; \mathcal{M}_{P^\perp_{mk}(S_u)}\big(\mathbf{1}_{mk}^\top\rho(\tilde{u} + \bullet)\,;\, P_{mk}^\perp\tilde{\mathbf{h}}_{\alpha,\sigma}, \frac{1}{r_1\nu_1}\big) \;, \\ M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu} \;&:=\; \mathcal{M}_{P_{mk}(S_u)}\big(M^\perp_{\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu}\,;\, \frac{1}{r_2\nu_2}\mathbf{y} - P_{mk}\tilde{\mathbf{h}}_{\alpha,\sigma}\,,\, r_2\nu_2\big) \;. \end{aligned}$$

Then (122) can be expressed as

$$\begin{aligned} & \min_{\tilde{u}\in P_{mk}(S_u)} \frac{1}{mk}M^\perp_{\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu}(\tilde{u}) + \frac{r_2\nu_2}{2mk}\Big\|P_{mk}\big(\tilde{u} - \frac{1}{r_1\nu_1}\mathbf{y} - \tilde{\mathbf{h}}_{\alpha,\sigma}\big)\Big\|^2 - \frac{1}{2r_2\nu_2 mk}\Big\|P_{mk}\mathbf{y}\Big\|^2 - \frac{1}{mk}\mathbf{y}^\top P_{mk}\tilde{\mathbf{h}}_{\alpha,\sigma} \\ = \;& \frac{1}{mk}M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu} - \frac{1}{2r_2\nu_2 mk}\|\mathbf{y}\|^2 - \frac{1}{mk}\mathbf{y}^\top\tilde{\mathbf{h}}_{\alpha,\sigma} \;, \end{aligned} \tag{123}$$

where we have used $P_{mk}\mathbf{y} = \mathbf{y}$ again in the last line. Substituting both (121) and (123) into (118) yields

$$\begin{aligned} \min_{\substack{\alpha\in S^\alpha \\ (\sigma_1,\sigma_2)\in S_{\sigma_1}\times S_{\sigma_2} \\ \nu_1,\nu_2\geq 0}} \;\; \max_{\substack{(r_1,r_2)\in S_{r_1}\times S_{r_2} \\ \tau_1,\tau_2\geq 0 \\ \theta\in\mathbb{R}}} \;\; & - \frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} + \alpha\theta\kappa_*^2 - \frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2} \\ & + M_{\mathbf{g},\sigma,\tau,\theta} - \frac{1}{4}\Big\|(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}\big(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)\big)\Big\|^2 \\ & + \frac{1}{mk}M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu} - \frac{1}{2r_2\nu_2 mk}\|\mathbf{y}\|^2 - \frac{1}{mk}\mathbf{y}^\top\tilde{\mathbf{h}}_{\alpha,\sigma} \;, \end{aligned}$$

which equals $R^{\mathrm{SO}}_{S,S_u,S_v}$.

Justifying the flipping of the minima and maxima. To conclude, we need to justify the flipping of min-max in the analysis of the auxiliary optimization above. The same argument has been performed for the logistic loss in the isotropic, unaugmented case in Salehi et al. (2019) and in more details for general losses in Thrampoulidis et al. (2018). For completeness, we repeat the arguments of the proof of Lemma A.3 of Thrampoulidis et al. (2018) in our context to illustrate how the non-asymptotic inequalities arise in our result for one particular flipping, and refer readers to Appendix A of Thrampoulidis et al. (2018) for more details in the general setup.

We now consider the flipping of $\min_{\beta\in S}$ and $\max_{(r_1,r_2)\in S_{r_1}\times S_{r_2}}$ in (110). First define the loss function $\mathcal{L}_1(\beta,u,\mu,w,r_1,r_2)$ as in (110) and denote the risk at (110) by

$$\mathcal{R}_1 \;:=\; \min_{\substack{\beta\in S\\ u\in\mathcal{S}_u\\ \mu\in\mathcal{S}}} \;\max_{\substack{w\in\mathbb{R}^p\\ (r_1,r_2)\in S_{r_1}\times S_{r_2}}} \;\mathcal{L}_1(\beta,u,\mu,w,r_1,r_2)\;.$$

For convenience, we also abbreviate the dependence on random variables in

$$R^{\mathrm{PO}}_{S,S_u,S_v} \;=R^{\mathrm{PO}}_{S,S_u,S_v}(\mathbf{G}\Sigma_o^{1/2}\beta^*,\mathbf{G}^{\Phi}\Sigma^{1/2})\;, \qquad L^{\mathrm{PO}}_{\beta,u,v} \;=L^{\mathrm{PO}}_{\beta,u,v}(\mathbf{G}\Sigma_o^{1/2}\beta^*,\mathbf{G}^{\Phi}\Sigma^{1/2})\;,$$
$$L^{\mathrm{AO}}_{\beta,u,v} \;=L^{\mathrm{AO}}_{\beta,u,v}(\mathbf{y},\mathbf{G}^{\Phi}P_*,\mathbf{g},\mathbf{h})\;.$$

By the computation of the auxiliary formulation up to (110), and by Lemma 48, we can apply Theorem 5(i) and (ii) to obtain that

$$\mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v}\;\leq\;c)\;\leq 4\,\mathbb{P}(\mathcal{R}_1\;\leq\;c) \qquad\text{and}\qquad \mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v}\;\geq\;c)\;\leq 4\,\mathbb{P}(\mathcal{R}_1\;\geq\;c) \tag{124}$$

for all $c\in\mathbb{R}$. Now define

$$\mathcal{R}_1' \;:=\; \max_{(r_1,r_2)\in S_{r_1}\times S_{r_2}} \;\min_{\substack{\beta\in S\\ u\in\mathcal{S}_u\\ \mu\in\mathcal{S}}} \;\max_{w\in\mathbb{R}^p} \mathcal{L}_1(\beta,u,\mu,w,r_1,r_2)\;.$$

By the min-max inequality (Rockafellar (1970), Lemma 36.1), we have $\mathcal{R}_1'\leq\mathcal{R}_1$ and therefore

$$\mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v}\;\leq\;c)\;\leq 4\,\mathbb{P}(\mathcal{R}_1\;\leq\;c)\;\leq\;4\,\mathbb{P}(\mathcal{R}_1'\;\leq\;c)\;. \tag{125}$$

To relate $\{R^{\mathrm{PO}}_{S,S_u,S_v}\geq c\}$ to $\{\mathcal{R}_1'\geq c\}$, we apply the min-max theorem (Rockafellar (1970), Corollary 37.3.2) to obtain that

$$R^{\mathrm{PO}}_{S,S_u,S_v} \;=\; \min_{\substack{\beta\in S\\ u\in S_u}}\max_{v\in S_v} L^{\mathrm{PO}}_{\beta,u,v} \;=\; \max_{v\in S_v}\min_{\substack{\beta\in S\\ u\in S_u}} L^{\mathrm{PO}}_{\beta,u,v}\;,$$

and applying Theorem 5 gives

$$\mathbb{P}(R^{\mathrm{PO}}_{S,S_u,S_v}\geq c)\;\leq\;4\,\mathbb{P}\Big(\max_{v\in S_v}\min_{\substack{\beta\in S\\ u\in S_u}} L^{\mathrm{AO}}_{\beta,u,v}\;\geq\;c\Big)\,. \tag{126}$$

Now recall that

$$S_{r_1}\;:=\Big\{\frac{1}{\sqrt{mk}}\|P_{mk}v\|\,\Big|\,v\in S_v\Big\} \qquad\text{and}\qquad S_{r_2}\;:=\Big\{\frac{1}{\sqrt{mk}}\|P^{\perp}_{mk}v\|\,\Big|\,v\in S_v\Big\}\,.$$

Defining $\tilde{S}_v(r_1, r_2) \coloneqq \{v \in S_v \,|\, \frac{1}{\sqrt{mk}} \|P_{mk} v\| = r_1 \,,\, \frac{1}{\sqrt{mk}} \|P^{\perp}_{mk} v\| = r_2\}$, we can rewrite

$$\begin{aligned} \max_{v \in S_v} \min_{\substack{\beta \in S \\ u \in S_u}} L^{\text{AO}}_{\beta,u,v} &= \max_{(r_1,r_2) \in S_{r_1} \times S_{r_2}} \max_{\tilde{v}_1 \in \tilde{S}_v(r_1,r_2)} \min_{\substack{\beta \in S \\ u \in S_u}} L^{\text{AO}}_{\beta,u,\tilde{v}} \\ &\overset{(a)}{\leq} \max_{(r_1,r_2) \in S_{r_1} \times S_{r_2}} \min_{\substack{\beta \in S \\ u \in S_u}} \max_{\tilde{v} \in \tilde{S}_v(r_1,r_2)} L^{\text{AO}}_{\beta,u,\tilde{v}} \overset{(b)}{=} \mathcal{R}'_1 \,, \end{aligned}$$

where we have applied the min-max inequality (Rockafellar (1970), Lemma 36.1) in (a) followed by the same computation up to (110) to maximize the loss over $\tilde{v}$. Combining this with (126) gives

$$\mathbb{P}(R^{\text{PO}}_{S,S_u,S_v} \geq c) \;\leq\; 4\,\mathbb{P}\Big(\max_{v \in S_v} \min_{\substack{\beta \in S \\ u \in S_u}} L^{\text{AO}}_{\beta,u,v} \;\geq\; c \Big) \;\leq\; 4\,\mathbb{P}(\mathcal{R}'_1 \;\geq\; c)\;.$$

Together with (125), this shows that $\mathcal{R}'_1$ is equivalent to $\mathcal{R}_1$ in the sense that the CGMT inequalities of (124) hold also with $\mathcal{R}_1$ replaced by $\mathcal{R}'_1$, therefore justifying the flipping of the minimum and the maximum. The remaining flipping of minimum and maximum over compact sets hold for the same reason, and any flipping that involves the Lagrange multiplier $w \in \mathbb{R}^p$ in (110) can be done in a similar manner by introducing the Lagrange multiplier directly to the (PO). This proves the first statement that for all $c \in \mathbb{R}$,

$$\mathbb{P}(R^{\text{PO}}_{S,S_u,S_v} \leq c) \;\leq 4\,\mathbb{P}(R^{\text{SO}}_{S,S_u,S_v} \;\leq\; c) \qquad \text{and} \qquad \mathbb{P}(R^{\text{PO}}_{S,S_u,S_v} \geq c) \;\leq 4\,\mathbb{P}(R^{\text{SO}}_{S,S_u,S_v} \;\geq\; c)\;.$$

Partial statement when S is replaced by $S_{c,\epsilon}$. When the optimization is over $S_{c,\epsilon}$, which is no longer compact, we cannot apply the min-max theorem for flipping min and max that involve $S_{c,\epsilon}$. However, notice that this change only affects optimizations over β, α, σ_1 and σ_2. For the optimization over β, we have shown in (125) that for the desired partial bound, the flipping of min and max does not require the min-max theorem. For the optimizations over α, σ_1 and σ_2, notice that the new domains of optimization for each of these variables are

$$\Big\{ \tfrac{v(\beta^*)^\intercal \Sigma^{1/2} \beta}{\sqrt{p}\,\kappa_*^2} \,\Big|\, \beta \in S \Big\}\,, \quad \{ \|(I_p - \Sigma_*) P_*^{\perp} \Sigma^{1/2} \beta\| \,\big|\, \beta \in S \}\,, \quad \{ \|\Sigma_* P_*^{\perp} \Sigma^{1/2} \beta\| \,\big|\, \beta \in S \}\,,$$

which are the 1d images of continuous functions on $\mathbb{R}^p$. Since $S_{c,\epsilon}$ is connected, the above sets are connected and therefore convex since they are one-dimensional. Therefore the replacement of S by $S_{c,\epsilon}$ does not affect the application of min-max theorem that concerns α, σ_1 and σ_2. This proves the partial upper bound analogous to (125): For all $c \in \mathbb{R}$,

$$\mathbb{P}(R^{\text{PO}}_{S_{c,\epsilon},S_u,S_v} \leq c) \;\leq 4\,\mathbb{P}(R^{\text{SO}}_{S_{c,\epsilon},S_u,S_v} \;\leq\; c)\;.$$

$\blacksquare$

N.3. Proof of Lemma 50: Equivalence between (SO) and (DO)

It is convenient to restate the optimization (SO):

$$\min_{\substack{\alpha\in S^\alpha\\(\sigma_1,\sigma_2)\in S_{\sigma_1}\times S_{\sigma_2}\\\nu_1,\nu_2\geq 0}}\ \max_{\substack{(r_1,r_2)\in S_{r_1}\times S_{r_2}\\\tau_1,\tau_2\geq 0\\\theta\in\mathbb{R}}}\ -\frac{\sigma_1}{2\tau_1}-\frac{\sigma_2}{2\tau_2}+\frac{r_1}{2\nu_1}+\frac{r_2}{2\nu_2}+\alpha\theta\kappa_*^2-\frac{\alpha^2\kappa_*^2}{2\sigma_2\tau_2}$$
$$+M_{\mathbf{g},\sigma,\tau,\theta}-\frac{1}{4}\Big\|(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2$$
$$+\frac{1}{mk}M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu}-\frac{1}{2r_2\nu_2mk}\|\mathbf{y}\|^2-\frac{1}{mk}\mathbf{y}^\intercal\tilde{\mathbf{h}}_{\alpha,\sigma}\ ,$$

As with Salehi et al. (2019), we exploit the fact that the optimization is over finitely many one-dimensional variables. It therefore suffices to analyze the asymptotics of the loss function directly, as one may first approximate the minimization and maximization over $(\tilde{\alpha},\tilde{\theta})$ by a smooth function and then take the approximation error to zero as $m,p\to\infty$.

Compute terms involving $\mathbf{g}_1$ and $\mathbf{g}_2$. Recall that $\mathbf{g}_1$ and $\mathbf{g}_2$ are independent standard Gaussian $\mathbb{R}^p$ vectors, and that

$$\begin{aligned}\tilde{\mathbf{g}}&=-\frac{r_1+r_2}{\sqrt{mk}}(P_\Sigma-\Sigma_*)\mathbf{g}_1-\frac{r_2}{\sqrt{m}}\Sigma_*\mathbf{g}_2\ ,\\ \tilde{\Sigma}_{\sigma,\tau}&=\frac{1}{2\sigma_1\tau_1}(P_\Sigma-\Sigma_*)+\frac{1}{2\sigma_2\tau_2}\Sigma_*\ ,\\ M_{\mathbf{g},\sigma,\tau,\theta}&=\min_{\mu\in S}\frac{\lambda}{2n}\|P_\Sigma\mu\|_2^2+\Big\|\tilde{\Sigma}^{1/2}_{\sigma,\tau}(\Sigma^{1/2}\mu)-\frac{1}{2}(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2\\ &\quad+\frac{r_2}{\sqrt{m}}\mathbf{g}_2^\intercal P_*\Sigma^{1/2}\mu\ .\end{aligned}$$

We focus on handling $M_{\mathbf{g},\sigma,\tau,\theta}$. First note that $\frac{r_2}{\sqrt{n}}\mathbf{g}_2^\intercal P_*\Sigma^{1/2}\mu$ depends on μ through the scalar $v(\beta^*)\Sigma^{1/2}\mu$, so by a similar reasoning as above, we can apply the law of large numbers directly to this term and obtain that it converges to zero in probability. Using $o_\mathbb{P}(1)$ to denote terms that converge in probability to zero, we then have

$$\begin{aligned}M_{\mathbf{g},\sigma,\tau,\theta}&=o_\mathbb{P}(1)+\min_{\mu\in S}\frac{\lambda}{2m}\|P_\Sigma\mu\|_2^2+\Big\|\tilde{\Sigma}^{1/2}_{\sigma,\tau}(\Sigma^{1/2}\mu)-\frac{1}{2}(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2\\ &=\tilde{M}_{\mathbf{g},\sigma,\tau,\theta}+\frac{1}{4}\Big\|(\tilde{\Sigma}^\dagger_{\sigma,\tau})^{1/2}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2+o_\mathbb{P}(1)\ ,\end{aligned}\tag{127}$$

where

$$\tilde{M}_{\mathbf{g},\sigma,\tau,\theta}:=\min_{\mu\in S}(\Sigma^{1/2}\mu)^\intercal\Big(\frac{\lambda}{2n}\Sigma^\dagger+\tilde{\Sigma}_{\sigma,\tau}\Big)(\Sigma^{1/2}\mu)-(\Sigma^{1/2}\mu)^\intercal(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\ .$$

The second term of (127) cancels with the other $(\mathbf{g}_1,\mathbf{g}_2)$-dependent term in the overall loss, so the only remaining $(\mathbf{g}_1,\mathbf{g}_2)$-dependent term is $\tilde{M}_{\mathbf{g},\sigma,\tau,\theta}$. By a completion of squares, we obtain

$$\begin{aligned}\tilde{M}_{\mathbf{g},\sigma,\tau,\theta}&=\min_{\mu\in S}\Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger+\tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}(\Sigma^{1/2}\mu)-\frac{1}{2}\Big(\Big(\frac{\lambda}{2m}\Sigma^\dagger+\tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger\Big)^{1/2}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2\\ &\quad-\frac{1}{4}(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))^\intercal\Big(\frac{\lambda}{2m}\Sigma^\dagger+\tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}}+\frac{\theta}{\sqrt{p}}v(\beta^*))\ .\end{aligned}$$

The second term does not involve μ, so we first seek to take a limit of this term. Recall from (120) that P_*, $P_\Sigma - \Sigma_*$ and $\Sigma_* - P_*$ are projections onto mutually orthogonal subspaces and that $P_\Sigma \Sigma_* = \Sigma_*$, $P_\Sigma P_* = P_*$. We can then express

$$\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger = \Big(\frac{\lambda}{2m}\Sigma^\dagger + \frac{1}{2\sigma_1\tau_1}(P_\Sigma - \Sigma_*) + \frac{1}{2\sigma_2\tau_2}\Sigma_*\Big)^\dagger ,$$

which implies that

$$\begin{aligned}\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)) \;=\; & -\frac{2(r_1+r_2)\sigma_1\tau_1}{\sqrt{mk}}\Big(\frac{2\sigma_1\tau_1\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger(P_\Sigma - \Sigma_*)\mathbf{g}_1 \\ & - \frac{2r_2\sigma_2\tau_2}{\sqrt{m}}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger \Sigma_*\mathbf{g}_2 \\ & + \frac{2\sigma_2\tau_2\theta}{\sqrt{p}}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger v(\beta^*) \\ & + O(\|m^{-1/2}P_*\mathbf{g}_1\| + \|m^{-1/2}P_*\mathbf{g}_2\|) \,. \end{aligned} \tag{128}$$

Also notice that any term linear in $\mathbf{g}_1$ or $\mathbf{g}_2$ has expectation zero, which vanishes. Recall also that $v(\beta^*) = \sqrt{p}\Sigma_*\Sigma_o^{1/2}\beta^*$ and $\kappa_* = \|\Sigma_*\Sigma_o^{1/2}\beta^*\|$. Computing the inverse along each orthogonal subspace explicitly and taking a limit with $m, p \to \infty$ and $p/n = p/(mk) \to \kappa$, we obtain

$$\begin{aligned} & -\frac{1}{4}(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))^\intercal\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*)) \\ & \quad \xrightarrow{\mathbb{P}} -\frac{(r_1+r_2)^2\sigma_1\tau_1}{2k}\bar{\chi}_{11}^{\sigma,\tau} - \frac{r_2^2\sigma_2\tau_2}{2}\bar{\chi}_{12}^{\sigma,\tau} - \frac{\theta^2\bar{\kappa}_*^2\sigma_2\tau_2}{2}\bar{\chi}_{13}^{\sigma,\tau} \;=\; -\bar{\chi}_1^{r,\theta,\sigma,\tau} \,. \end{aligned}$$

where we have recalled that $P_* = v(\beta^*)v(\beta^*)^\intercal/(p\kappa_*^2)$, $\bar{\kappa}_* = \lim \kappa_*$ and

$$\begin{aligned} \bar{\chi}_{11}^{\sigma,\tau} &:= \lim \frac{\mathrm{Tr}\,((\frac{2\sigma_1\tau_1\lambda}{2m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*))}{m} \,, \\ \bar{\chi}_{12}^{\sigma,\tau} &:= \lim \frac{\mathrm{Tr}\,((\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p)^\dagger\Sigma_*)}{m} \,, \\ \bar{\chi}_{13}^{\sigma,\tau} &:= \lim \mathrm{Tr}\Big(\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big) \,. \end{aligned}$$

To address the minimization over $\mu \in S$, notice that the only difference between the two choices of S are via the restriction on $\mu^\intercal\Sigma_{\mathrm{new}}\mu$. Recall that the two different choices of S differs only in $\beta^\intercal\Sigma_{\mathrm{new}}\beta$. Let $P_{\Sigma_{\mathrm{new}}}$ be the projection onto the positive eigenspace of $P_{\Sigma_{\mathrm{new}}}$ and $P^\perp_{\Sigma_{\mathrm{new}}} = I_p - P_{\Sigma_{\mathrm{new}}}$. Then we can rewrite the minimization as

$$\min_{\substack{\mu \in P_{\Sigma_{\mathrm{new}}}(S) \\ \mu' \in P^\perp_{\Sigma_{\mathrm{new}}}(S)}} \Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}\Big(\mu + \mu' - \frac{1}{2}(\Sigma^\dagger)^{1/2}\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\Big)\Big\|^2 \,.$$

With either choice of S, μ' can be chosen freely within $P^\perp_{\Sigma_{\mathrm{new}}}(\mathbb{R}^p)$ so long as $\|\mu'\|_2 = O(\sqrt{p})$. Minimizing over μ' first and noting that $P_{\Sigma_{\mathrm{new}}} = (\Sigma^\dagger_{\mathrm{new}})^{1/2}\Sigma^{1/2}_{\mathrm{new}}$, we obtain

$$\min_{\mu \in P_{\Sigma_{\mathrm{new}}}(S)} \Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}(\Sigma^\dagger_{\mathrm{new}})^{1/2}\Big(\Sigma^{1/2}_{\mathrm{new}}\mu - \frac{1}{2}\Sigma^{1/2}_{\mathrm{new}}(\Sigma^\dagger)^{1/2}\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\Big)\Big\|^2 \,.$$

Setting $\Sigma_{\text{new}}^{1/2}\mu$ in the direction of minimization, we have that for some $c(\mu) \in \mathbb{R}$,

$$\Sigma_{\text{new}}^{1/2}\mu \;=\; c(\mu)\mathbf{g}'\;, \qquad \mathbf{g}' \;:=\; \tfrac{1}{2}\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\;,$$

which allows us to rewrite

$$\frac{\Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}(\Sigma_{\text{new}}^\dagger)^{1/2}\mathbf{g}'\Big\|^2}{\|\mathbf{g}'\|^2} \;\times\; \min_{\mu \in P_{\Sigma_{\text{new}}}(S)}\;(c(\mu)\|\mathbf{g}'\| - \|\mathbf{g}'\|)^2\;.$$

This is now an optimization over a scalar, so we can again take the limit inside the minimization. We proceed to compute the limits of the two norms involving $\mathbf{g}'$. Recycling the computation in (128), we have

$$\begin{aligned}\|\mathbf{g}'\|^2 \;&=\; \frac{(r_1+r_2)^2\sigma_1^2\tau_1^2}{mk}\,\Big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{2\sigma_1\tau_1\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger(P_\Sigma - \Sigma_*)\mathbf{g}_1\Big\|^2\\ &\quad+\; \frac{r_2^2\sigma_2^2\tau_2^2}{n}\,\Big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger P_*\mathbf{g}_2\Big\|^2\\ &\quad+\; \frac{\theta^2\sigma_2^2\tau_2^2}{p}\Big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger v(\beta^*)\Big\|^2 + o_{\mathbb{P}}(1)\\ &\xrightarrow{\mathbb{P}}\; \frac{(r_1+r_2)^2\sigma_1^2\tau_1^2}{k}\bar{\chi}_{21}^{\sigma,\tau} + r_2^2\sigma_2^2\tau_2^2\,\bar{\chi}_{22}^{\sigma,\tau} + \theta^2\bar{\kappa}_*^2\sigma_2^2\tau_2^2\,\bar{\chi}_{23}^{\sigma,\tau} \;=\; \bar{\chi}_2^{r,\theta,\sigma,\tau}\;,\end{aligned}$$

where we have denoted

$$\begin{aligned}\bar{\chi}_{21}^{\sigma,\tau} \;&:=\; \lim\frac{\big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}(\frac{2\sigma_1\tau_1\lambda}{2m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*)\big\|^2}{m}\;,\\ \bar{\chi}_{22}^{\sigma,\tau} \;&:=\; \lim\frac{\big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\big)^\dagger\Sigma_*\big\|^2}{m}\;,\\ \bar{\chi}_{23}^{\sigma,\tau} \;&:=\; \lim\;\Big\|\Sigma_{\text{new}}^{1/2}(\Sigma^\dagger)^{1/2}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big\|^2\;.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}(\Sigma_{\text{new}}^\dagger)^{1/2}\mathbf{g}'\Big\|^2\\ &= \Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}P_{\Sigma_{\text{new}}}(\Sigma^\dagger)^{1/2}\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^\dagger(\tilde{\mathbf{g}} + \frac{\theta}{\sqrt{p}}v(\beta^*))\Big\|^2\\ &\xrightarrow{\mathbb{P}}\; \frac{(r_1+r_2)^2\sigma_1^2\tau_1^2}{k}\bar{\chi}_{31}^{\sigma,\tau} + r_2^2\sigma_2^2\tau_2^2\,\bar{\chi}_{32}^{\sigma,\tau} + \theta^2\bar{\kappa}_*^2\sigma_2^2\tau_2^2\,\bar{\chi}_{33}^{\sigma,\tau} \;=\; \bar{\chi}_3^{r,\theta,\sigma,\tau}\;,\end{aligned}$$

where we used

$$\begin{aligned}\bar{\chi}_{31}^{\sigma,\tau} \;&:=\; \lim\frac{\big\|(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau})^{1/2}\Sigma^{1/2}P_{\Sigma_{\text{new}}}(\Sigma^\dagger)^{1/2}(\frac{2\sigma_1\tau_1\lambda}{2m}\Sigma^\dagger + I_p)^\dagger(P_\Sigma - \Sigma_*)\big\|^2}{m}\;,\\ \bar{\chi}_{32}^{\sigma,\tau} \;&:=\; \lim\frac{\big\|(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau})^{1/2}\Sigma^{1/2}P_{\Sigma_{\text{new}}}(\Sigma^\dagger)^{1/2}\big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\big)^\dagger\Sigma_*\big\|^2}{m}\;,\\ \bar{\chi}_{33}^{\sigma,\tau} \;&:=\; \lim\;\Big\|\Big(\frac{\lambda}{2m}\Sigma^\dagger + \tilde{\Sigma}_{\sigma,\tau}\Big)^{1/2}\Sigma^{1/2}P_{\Sigma_{\text{new}}}(\Sigma^\dagger)^{1/2}\Big(\frac{2\sigma_2\tau_2\lambda}{2m}\Sigma^\dagger + I_p\Big)^\dagger P_*\Big\|^2\;.\end{aligned}$$

Combining the calculations above, we obtain that

$$\tilde{M}_{\mathbf{g},\sigma,\tau,\theta} \;\xrightarrow{\mathbb{P}}\; -\bar{\chi}_1^{r,\theta,\sigma,\tau} + \frac{\bar{\chi}_3^{r,\theta,\sigma,\tau}}{\bar{\chi}_2^{r,\theta,\sigma,\tau}} \;\times\; \min_{\mu \in P_{\Sigma_{\text{new}}}(S)}\Big(c(\mu)\sqrt{\bar{\chi}_2^{r,\theta,\sigma,\tau}} - \sqrt{\bar{\chi}_2^{r,\theta,\sigma,\tau}}\Big)^2\;.$$

Notice that $\|\Sigma_{\text{new}}^{1/2}\mu\| = c(\mu)\|\mathbf{g}'\| \xrightarrow{\mathbb{P}} c(\mu)\sqrt{\bar{\chi}_2^{r,\theta,\sigma,\tau}}$, and recall that our two choices of S only differs through $\|\Sigma_{\text{new}}^{1/2}\mu\| - (\bar{\chi}_2^{\bar{r},\bar{\theta},\bar{\sigma},\bar{\tau}})^{1/2}$, where $\bar{r} = (\bar{r}_1, \bar{r}_2)$, $\bar{\sigma} = (\bar{\sigma}_1, \bar{\sigma}_2)$, $\bar{\theta}$ and $\bar{\tau} = (\bar{\tau}_1, \bar{\tau}_2)$ are the optimal solutions to (DO). This implies that

$$\tilde{M}_{\mathbf{g},\sigma,\tau,\theta} \xrightarrow{\mathbb{P}} -\bar{\chi}_1^{r,\theta,\sigma,\tau} + \epsilon_S^2 \frac{\bar{\chi}_3^{r,\theta,\sigma,\tau}}{\bar{\chi}_2^{r,\theta,\sigma,\tau}} ,$$

where $\epsilon_S = 0$ for $S = \mathcal{S}_p$ and $\epsilon_S = \epsilon$ for $S = \mathcal{S}_\epsilon^c$. Substituting this back into the overall optimization, we can approximate (SO) in distribution by

$$\begin{aligned} \min_{\substack{\alpha\in S^\alpha \\ (\sigma_1,\sigma_2)\in S_{\sigma_1}\times S_{\sigma_2} \\ \nu_1,\nu_2\geq 0}} \quad \max_{\substack{(r_1,r_2)\in S_{r_1}\times S_{r_2} \\ \tau_1,\tau_2\geq 0 \\ \theta\in\mathbb{R}}} \quad & -\frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} + \alpha\theta\bar{\kappa}_*^2 - \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2} - \bar{\chi}_1^{r,\theta,\sigma,\tau} + \epsilon_S^2 \frac{\bar{\chi}_3^{r,\theta,\sigma,\tau}}{\bar{\chi}_2^{r,\theta,\sigma,\tau}} \\ & + \frac{1}{mk} M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu} - \frac{1}{2r_2\nu_2 mk}\|\mathbf{y}\|^2 - \frac{1}{mk}\mathbf{y}^\intercal\tilde{\mathbf{h}}_{\alpha,\sigma} . \end{aligned}$$

Compute terms involving y and $\tilde{\mathbf{h}}_{\alpha,\sigma}$. Recall that $\tilde{\mathbf{h}}_{\alpha,\sigma} = \kappa_*\alpha\mathbf{q} - \sigma_1\mathbf{h}_1 - \frac{\sigma_2}{\sqrt{k}} J_{mk}\mathbf{h}_2$, where $\mathbf{q} = \mathbf{q}(\mathbf{G}^\Phi) = \frac{1}{\kappa_*\sqrt{p}}\mathbf{G}^\Phi v(\beta^*)$, and $\mathbf{h}_1$ and $\mathbf{h}_2$ are i.i.d. standard $\mathbb{R}^{mk}$ Gaussians independent of $\mathbf{q}(\mathbf{G}^\Phi)$ and $\mathbf{y} = \mathbf{y}(\mathbf{G})$. Also recall that $\mathbf{y} = \mathbf{y}P_{mk} = \frac{1}{k}\mathbf{y}J_{mk}$. We can then express the last two terms of the loss above as

$$-\frac{1}{2r_2\nu_2 mk}\|\mathbf{y}\|^2 - \frac{1}{mk}\mathbf{y}^\intercal\tilde{\mathbf{h}}_{\alpha,\sigma} = -\frac{1}{2r_2\nu_2 mk}\|\mathbf{y}\|^2 - \frac{\kappa_*\alpha}{mk}\mathbf{y}^\intercal\mathbf{q} + \frac{\sigma_1}{mk}\mathbf{y}^\intercal\mathbf{h}_1 + \frac{\sigma_2}{m\sqrt{k}}\mathbf{y}^\intercal P_{mk}\mathbf{h}_2 . \tag{129}$$

Since $\mathbf{h}_1$ and $\mathbf{h}_2$ are zero-mean and $\mathbf{y}$ is coordinate-wise bounded by one, by the weak law of large numbers,

$$\frac{1}{mk}\mathbf{y}^\intercal\mathbf{h}_1 \xrightarrow{\mathbb{P}} 0 \qquad \text{and} \qquad \frac{1}{m\sqrt{k}}\mathbf{y}^\intercal P_{mk}\mathbf{h}_2 \xrightarrow{\mathbb{P}} 0 . \tag{130}$$

To handle the first two terms, recall that

$$\mathbf{y} = (\underbrace{y_1(\Sigma_o^{1/2}G_1), \ldots, y_1(\Sigma_o^{1/2}G_1)}_{\text{repeated } k \text{ times}}, \ \ldots, \ \underbrace{y_n(\Sigma_o^{1/2}G_n), \ldots, y_n(\Sigma_o^{1/2}G_n)}_{\text{repeated } k \text{ times}})^\intercal ,$$

where $y_i(\Sigma_o^{1/2}G_i)$'s are i.i.d. by definition. Therefore by the weak law of large numbers,

$$\frac{1}{mk}\|\mathbf{y}\|^2 = \frac{1}{m}\sum_{i=1}^m y_i(\Sigma_o^{1/2}G_i) \xrightarrow{\mathbb{P}} \mathbb{E}[y_1(\Sigma_o^{1/2}G_1)] = \frac{1}{2} . \tag{131}$$

In the last equality, we recall that

$$\mathbb{P}(y_1(\Sigma_o^{1/2}G_1) = 1 \,\big|\, G_1) = \sigma(G_1^\intercal\Sigma_o^{1/2}\beta^*) .$$

$y_i \in \{0, 1\}$ is a logistic variable evaluated at a random input $(\Sigma_o^{1/2}G_1)^\intercal\beta^*$ that is symmetric about zero. On the other hand, recalling that $v(\beta^*) = \sqrt{p}\Sigma_*\Sigma_o^{1/2}\beta^*$,

$$\begin{aligned} \frac{1}{mk}\mathbf{y}^\intercal\mathbf{q} &= \frac{1}{mk}\sum_{i\leq m}\sum_{j\leq k} y_i(\Sigma_o^{1/2}G_i)\,\frac{1}{\kappa_*\sqrt{p}}(G_{ij}^\Phi)^\intercal v(\beta^*) \\ &= \frac{1}{\kappa_*}\frac{1}{mk}\sum_{i\leq m}\sum_{j\leq k} y_i(\Sigma_o^{1/2}G_i)\,(G_{ij}^\Phi)^\intercal\Sigma_*\Sigma_o^{1/2}\beta^* . \end{aligned} \tag{132}$$

Notice that each $y_i(\Sigma_o^{1/2} G_i)$ depends on G_i only through $G_i^\intercal \Sigma_o^{1/2} \beta^*$, so G_i and G_{ij}^Φ appear in each summand only via the Gaussian vector

$$\begin{pmatrix} G_i^\intercal \Sigma_o^{1/2} \beta^* \\ \big(G_{i1}^\Phi\big)^\intercal \Sigma_* \Sigma_o^{1/2} \beta^* \\ \vdots \\ \big(G_{ik}^\Phi\big)^\intercal \Sigma_* \Sigma_o^{1/2} \beta^* \end{pmatrix} \sim \mathcal{N}\Bigg(0\,, \begin{pmatrix} (\beta^*)^\intercal \Sigma_o \beta^* & (\beta^*)^\intercal \Sigma_o^{1/2} \Sigma_* \Sigma_o^{1/2} \beta^* & \dots \\ (\beta^*)^\intercal \Sigma_o^{1/2} \Sigma_* \Sigma_o^{1/2} \beta^* & (\beta^*)^\intercal \Sigma_o^{1/2} \Sigma_* \Sigma_o^{1/2} \beta^* & \\ \vdots & & \ddots \end{pmatrix}\Bigg) \xrightarrow{d} \begin{pmatrix} \bar\kappa_o \bar Z_0 + \bar\kappa_* \bar Z_1 \\ \bar\kappa_* \bar Z_1 \\ \vdots \\ \bar\kappa_* \bar Z_1 \end{pmatrix},$$

where we recall that by Assumption 11,

$$\begin{aligned} \Sigma_o^{1/2}\, \mathrm{Cov}[G_i, G_{i1}^\Phi]\, \Sigma_* \Sigma_o^{1/2} &= \Sigma_o^{1/2} (\Sigma_*)^2 \Sigma_o^{1/2} \;=\; \Sigma_o^{1/2} \Sigma_* \Sigma_o^{1/2}\,, \\ \Sigma_o^{1/2}\, \Sigma_* \mathrm{Cov}[G_{i1}^\Phi, G_{i2}^\Phi]\, \Sigma_* \Sigma_o^{1/2} &= \Sigma_o^{1/2} (\Sigma_*)^2 \Sigma_o^{1/2} \;=\; \Sigma_o^{1/2} \Sigma_* \Sigma_o^{1/2}\,, \end{aligned}$$

$\bar Z_0$ and $\bar Z_1$ are two i.i.d. standard normals and

$$\bar\kappa_* \;\coloneqq\; \lim_{p\to\infty} \kappa_* \;=\; \lim_{p\to\infty} \|\Sigma_* \Sigma_o^{1/2} \beta^*\|\,, \qquad \bar\kappa_o \;\coloneqq\; \lim_{p\to\infty} \|(I_p - \Sigma_*) \Sigma_o^{1/2} \beta^*\|\,.$$

Then by the law of large numbers, we have

$$\frac{\kappa_*}{mk} \mathbf{y}^\intercal \mathbf{q} \;\xrightarrow{\mathbb{P}}\; \mathbb{E}[\sigma(\bar\kappa_o \bar Z_0 + \bar\kappa_* \bar Z_1) \bar\kappa_* \bar Z_1]\,. \tag{133}$$

Combining (129), (130), (131) and (133) gives

$$-\frac{1}{2 r_2 \nu_2 mk} \|\mathbf{y}\|^2 - \frac{1}{mk} \mathbf{y}^\intercal \tilde{\mathbf{h}}_{\alpha,\sigma} \;\xrightarrow{\mathbb{P}}\; -\frac{1}{4 r_2 \nu_2} - \alpha\, \mathbb{E}[\sigma(\bar\kappa_o \bar Z_0 + \bar\kappa_* \bar Z_1) \bar\kappa_* \bar Z_1]\,,$$

so the optimization can be approximated by

$$\begin{aligned} \min_{\substack{\alpha \in S^\alpha \\ (\sigma_1,\sigma_2) \in S_{\sigma_1} \times S_{\sigma_2} \\ \nu_1,\nu_2 \geq 0}} \; \max_{\substack{(r_1,r_2) \in S_{r_1} \times S_{r_2} \\ \tau_1,\tau_2 \geq 0 \\ \theta \in \mathbb{R}}} \; & - \frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} + \alpha\theta\bar\kappa_*^2 - \frac{\alpha^2 \bar\kappa_*^2}{2\sigma_2\tau_2} - \bar\chi_1^{r,\theta,\sigma,\tau} + \epsilon_S^2 \frac{\bar\chi_3^{r,\theta,\sigma,\tau}}{\bar\chi_2^{r,\theta,\sigma,\tau}} - \frac{1}{4 r_2 \nu_2} \\ & - \alpha\, \mathbb{E}[\sigma(\bar\kappa_o \bar Z_0 + \bar\kappa_* \bar Z_1) \bar\kappa_* \bar Z_1] + \frac{1}{mk} M_{\mathbf{y}, \tilde{\mathbf{h}}_{\alpha,\sigma}, r, \nu}\,. \end{aligned} \tag{134}$$

Computing the nested Moreau envelope. We are left with

$$\begin{aligned} \frac{1}{mk} M_{\mathbf{y}, \tilde{\mathbf{h}}_{\alpha,\sigma}, r, \nu} &= \min_{u_2 \in P_{mk}(S_u)} \frac{1}{mk} M^\perp_{\tilde{\mathbf{h}}_{\alpha,\sigma}, r, \nu}(u_2) + \frac{r_2 \nu_2}{2mk} \Big\| P_{mk}\big(u_2 - \frac{1}{r_2 \nu_2} \mathbf{y} - \tilde{\mathbf{h}}_{\alpha,\sigma}\big) \Big\|^2 \\ &= \min_{u_2 \in P_{mk}(S_u)} \; \min_{u_1 \in P_{mk}^\perp(S_u)} \; \frac{1}{mk} \mathbf{1}_{mk}^\intercal \rho(u_1 + u_2) + \frac{r_2 \nu_2}{2mk} \Big\| P_{mk}\big(u_2 - \frac{1}{r_2 \nu_2} \mathbf{y} - \tilde{\mathbf{h}}_{\alpha,\sigma}\big) \Big\|^2 \\ &\qquad + \frac{r_1 \nu_1}{2mk} \Big\| P_{mk}^\perp (u_1 - \tilde{\mathbf{h}}_{\alpha,\sigma}) \Big\|^2\,. \end{aligned}$$

Write u_{1ij} as the (i,j)-th coordinate of $u_1 \in \mathbb{R}^{mk}$ and similarly write u_{2ij} for that of u_2, q_{ij} for $\mathbf{q}$, h_{1ij} for $\mathbf{h}_1$ and h_{2ij} for $\mathbf{h}_2$. Recalling the definition of ρ, $P_{mk} = \frac{1}{k} J_{mk}$ and $P_{mk}^\perp = I_{mk} - P_{mk}$, we can re-express the loss above as

$$\tfrac{1}{mk} \textstyle\sum_{i,j=1}^k L_{ij}(u_1, u_2)\,,$$

where

$$
\begin{aligned}
L_{ij}(u_1, u_2) \ :=\ & \log(1 + e^{u_{1ij}+u_{2ij}}) \\
& + \tfrac{r_2\nu_2}{2}\Big(\tfrac{1}{k}\textstyle\sum_{j'=1}^{k}\big(u_{2ij'} - \tfrac{1}{r_2\nu_2}y_i(\Sigma_o^{1/2}G_i) - \kappa_*\alpha q_{ij'} + \sigma_1 h_{1ij'} + \tfrac{\sigma_2}{\sqrt{k}}\big(\textstyle\sum_{j''\leq k} h_{2ij''}\big)\big)\Big)^2 \\
& + \tfrac{r_1\nu_1}{2}(u_{1ij} - \kappa_*\alpha q_{ij} + \sigma_1 h_{1ij})^2 - \tfrac{r_1\nu_1}{2}\Big(\tfrac{1}{k}\textstyle\sum_{j'=1}^{k}(u_{1ij'} - \kappa_*\alpha q_{ij'} + \sigma_1 h_{1ij'})\Big)^2 .
\end{aligned}
$$

Consider the $\mathbb{R}^k$-valued vectors $u_{1i} = (u_{1i1}, \dots, u_{1ik})$ and $u_{2i} = (u_{2i1}, \dots, u_{2ik})$ for $1 \leq i \leq m$. Notice that the loss $L_{ij}(u_1, u_2) = \tilde{L}_{ij}(u_{1i}, u_{2i})$ only depends on u_1 and u_2 through u_{1i}, u_{2i}. This allows us to rewrite

$$
\frac{1}{mk}M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu} \ =\ \frac{1}{m}\sum_{i=1}^{m} \min_{u_{2i}\in(P_{mk}(S_u))_i}\ \min_{u_{1i}\in(P_{mk}^{\perp}(S_u))_i} \frac{1}{k}\sum_{j=1}^{k}\tilde{L}_{ij}(u_{1i}, u_{2i}) \,,
$$

where $(P_{mk}(S_u))_i$ and $(P_{mk}^{\perp}(S_u))_i$ are the corresponding subspaces in which u_{2i} and u_{1i} take values. Since S_u is closed under permutation of its m blocks of k coordinates, the m summands above are i.i.d., which allows us to apply a weak law of large numbers to the above average. Also note that the minima are over $\mathbb{R}^k$-valued vectors, which allows us again to take a limit with $p \to \infty$ inside the loss function. Using the computation of $y_{ij}q_{ij}$ via $\bar{Z}_0$ and $\bar{Z}_1$ in (132), we obtain that $M_{\mathbf{y},\tilde{\mathbf{h}}_{\alpha,\sigma},r,\nu}$ can be approximated by

$$
\begin{aligned}
\mathbb{E}\Big[\min_{\substack{u'\in(P_{mk}(S_u))_1 \\ u''\in(P_{mk}^{\perp}(S_u))_1}} \ & \frac{1}{k}\sum_{j=1}^{k}\log\Big(1 + e^{u'_j+u''_j}\Big) \\
& + \frac{r_2\nu_2}{2}\Big(\frac{1}{k}\sum_{j=1}^{k}\big(u'_j - \frac{1}{r_2\nu_2}\mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j + \sigma_2\bar{Z}_2\big)\Big)^2 \\
& + \frac{r_1\nu_1}{2}\Big(\frac{1}{k}\sum_{j=1}^{k}(u''_j - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j)^2 - \Big(\frac{1}{k}\sum_{j=1}^{k}(u''_j - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j)\Big)^2\Big)\Big] ,
\end{aligned}
$$

where $\eta_1, \dots, \eta_k$ and $\bar{Z}_2$ are i.i.d. standard normals and ε_1 is an independent Logistic(0, 1) variable.. Notice that $u' \in (P_{mk}(S_u))_1$ has equal entries, say u_0, and $u'' \in (P_{mk}^{\perp}(S_u))_1$ satisfies $\sum_{j=1}^{k} u''_j = 0$. Also recall the assumption that $\sup_{u\in S_u}\frac{\|u\|_2^2}{mk} \to \infty$. Setting $\tilde{u} = (u''_1 + u_0, \dots, u''_k + u_0)$, the above can be further approximated by

$$
\begin{aligned}
& \mathbb{E}\Big[\min_{\tilde{u}\in\mathbb{R}^k}\ \frac{1}{k}\sum_{j=1}^{k}\log\Big(1 + e^{\tilde{u}_j}\Big) \\
& \qquad + \frac{r_2\nu_2}{2}\Big(\ \frac{1}{k}\sum_{j\leq k}\Big(\tilde{u}_j - \frac{1}{r_2\nu_2}\mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j + \sigma_2\bar{Z}_2\Big)\Big)^2 \\
& \qquad + \frac{r_1\nu_1}{2}\Big(\frac{1}{k}\sum_{j=1}^{k}(\tilde{u}_j - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j)^2 - \Big(\frac{1}{k}\sum_{j=1}^{k}(\tilde{u}_j - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_1\eta_j)\Big)^2\Big)\Big] \\
= \ & \mathbb{E}\Big[\min_{\tilde{u}\in\mathbb{R}^k}\ \frac{1}{k}\mathbf{1}_k^{\intercal}\rho(\tilde{u}) + \frac{r_1\nu_1}{2k}\Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(\tilde{u} - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta)\Big\|^2 \\
& \qquad + \frac{r_2\nu_2}{2k}\Big\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(\tilde{u} - \frac{1}{r_2\nu_2}\mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\Big\|^2\Big] ,
\end{aligned}
$$

which equals $\bar{M}_\rho^{r,\nu,\alpha,\sigma}$. Substituting this into (134) while also applying the assumption that $\sup_{v \in S_v} \frac{\|v\|_2^2}{mk} \to \infty$, we obtain

$$\min_{\substack{\alpha \in S^\alpha \\ (\sigma_1,\sigma_2) \in S_{\sigma_1} \times S_{\sigma_2} \\ \nu_1,\nu_2 \geq 0}} \max_{\substack{(r_1,r_2) \in S_{r_1} \times S_{r_2} \\ \tau_1,\tau_2 \geq 0 \\ \theta \in \mathbb{R}}} -\frac{\sigma_1}{2\tau_1} - \frac{\sigma_2}{2\tau_2} + \frac{r_1}{2\nu_1} + \frac{r_2}{2\nu_2} + \alpha\theta\bar{\kappa}_*^2 - \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2}$$

$$-\bar{\chi}_1^{r,\theta,\sigma,\tau} + \epsilon_S^2 \frac{\bar{\chi}_3^{r,\theta,\sigma,\tau}}{\bar{\chi}_2^{r,\theta,\sigma,\tau}} - \frac{1}{4r_2\nu_2} - \alpha\,\mathbb{E}[\sigma(\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1)\bar{\kappa}_*\bar{Z}_1] + \bar{M}_\rho^{r,\nu,\alpha,\sigma}\ .$$

■

N.4. Proof of Lemma 51: (DO) to (EQs)

Since $S = \mathcal{S}_p$, we can ignore terms involving $\bar{\chi}_2^{r,\theta,\sigma,\tau}$ and $\bar{\chi}_3^{r,\theta,\sigma,\tau}$. Setting the first derivative of (DO) to zero with respect to each variable, we obtain

$$\begin{cases} 0 = \theta\bar{\kappa}_*^2 - \frac{\alpha\bar{\kappa}_*^2}{\sigma_2\tau_2} - \mathbb{E}[\sigma(\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1)\bar{\kappa}_*\bar{Z}_1] + \partial_\alpha \bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = -\frac{1}{2\tau_1} - \partial_{\sigma_1}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \partial_{\sigma_1}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = -\frac{1}{2\tau_2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2^2\tau_2} - \partial_{\sigma_2}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \partial_{\sigma_2}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = \frac{\sigma_1}{2\tau_1^2} - \partial_{\tau_1}\bar{\chi}_1^{r,\theta,\sigma,\tau}\ , \\ 0 = \frac{\sigma_2}{2\tau_2^2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2^2} - \partial_{\tau_2}\bar{\chi}_1^{r,\theta,\sigma,\tau}\ , \\ 0 = -\frac{r_1}{2\nu_1^2} + \partial_{\nu_1}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = -\frac{r_2}{2\nu_2^2} + \frac{1}{4r_2\nu_2^2} + \partial_{\nu_2}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = \frac{1}{2\nu_1} - \partial_{r_1}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \partial_{r_1}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = \frac{1}{2\nu_2} + \frac{1}{4r_2^2\nu_2} - \partial_{r_2}\bar{\chi}_1^{r,\theta,\sigma,\tau} + \partial_{r_2}\bar{M}_\rho^{r,\nu,\alpha,\sigma}\ , \\ 0 = \alpha\bar{\kappa}_*^2 - \partial_\theta\bar{\chi}_1^{r,\theta,\sigma,\tau}\ . \end{cases} \tag{135}$$

The next step is to compute the derivatives of

$$\bar{M}_\rho^{r,\nu,\alpha,\sigma} := \mathbb{E}\Big[\min_{\tilde{u} \in \mathbb{R}^k} \frac{1}{k}\mathbf{1}_k^\top \rho(\tilde{u}) + \frac{r_1\nu_1}{2k}\left\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(\tilde{u} + \sigma_1\eta)\right\|^2$$

$$+ \frac{r_2\nu_2}{2k}\left\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(\tilde{u} - \frac{1}{r_2\nu_2}\mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\right\|^2\Big]\ .$$

Recall that we denote $u_{\bar{Z},\varepsilon_1,\eta}$ as the minimizer of the minimization inside the expectation. By the envelope theorem and noting that $\mathbb{E}[\mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\}] = \mathbb{E}[\sigma(\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1)]$, we have

$$\begin{aligned}
\partial_\alpha \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= -\tfrac{r_2 \nu_2 \bar{\kappa}_*}{k} \mathbb{E}\Big[\bar{Z}_1 \Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2 \nu_2} \sigma(\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1)\Big)\Big] + r_2 \nu_2 \alpha \bar{\kappa}_*^2 \;, \\
\partial_{\sigma_1} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{r_1 \nu_1}{k} \, \mathbb{E}\Big[\eta^\intercal \, (I_k - \frac{1}{k} \mathbf{1}_{k\times k}) \, u_{\bar{Z},\varepsilon_1,\eta}\Big] + \frac{r_1 \nu_1 \sigma_1 (k-1)}{k} + \tfrac{r_2 \nu_2}{k} \mathbb{E}\Big[\eta^\intercal \frac{1}{k} \mathbf{1}_{k\times k} u_{\bar{Z},\varepsilon_1,\eta}\Big] + \tfrac{r_2 \nu_2 \sigma_1}{k} \;, \\
\partial_{\sigma_2} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{r_2 \nu_2}{k} \mathbb{E}\Big[\bar{Z}_2 \mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta}\Big] + r_2 \nu_2 \sigma_2 \;, \\
\partial_{\nu_1} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{r_1}{2k} \, \mathbb{E}\Big[\Big\|(I_k - \frac{1}{k} \mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1 \eta)\Big\|^2\Big] \;, \\
\partial_{\nu_2} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{r_2}{2k} \mathbb{E}\Big[\Big\|\frac{1}{k} \mathbf{1}_{k\times k} \Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2 \nu_2} \mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\} \mathbf{1}_k - \alpha \bar{\kappa}_* \bar{Z}_1 \mathbf{1}_k + \sigma_1 \eta + \sigma_2 \bar{Z}_2 \mathbf{1}_k\Big)\Big\|^2\Big] \\
&\quad + \tfrac{1}{\nu_2 k} \mathbb{E}\Big[\mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\} \Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2 \nu_2} \sigma(\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1) - k \alpha \bar{\kappa}_* \bar{Z}_1\Big)\Big] \;, \\
\partial_{r_1} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{\nu_1}{2k} \, \mathbb{E}\Big[\Big\|(I_k - \frac{1}{k} \mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1 \eta)\Big\|^2\Big] \;, \\
\partial_{r_2} \bar{M}_\rho^{r,\nu,\alpha,\sigma} &= \tfrac{\nu_2}{2k} \mathbb{E}\Big[\Big\|\frac{1}{k} \mathbf{1}_{k\times k} \Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2 \nu_2} \mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\} \mathbf{1}_k - \alpha \bar{\kappa}_* \bar{Z}_1 \mathbf{1}_k + \sigma_1 \eta + \sigma_2 \bar{Z}_2 \mathbf{1}_k\Big)\Big\|^2\Big] \\
&\quad + \tfrac{1}{r_2 k} \mathbb{E}\Big[\mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\} \Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2 \nu_2} \sigma(\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1) - k \alpha \bar{\kappa}_* \bar{Z}_1\Big)\Big] \;.
\end{aligned}$$

Writing $\bar{Y} = \mathbb{I}_{\geq 0}\{\bar{\kappa}_o \bar{Z}_0 + \bar{\kappa}_* \bar{Z}_1 - \varepsilon_1\}$ and substituting the bounds above into the system of equations recovers (EQs).

N.5. Proofs for Appendix M.2

N.5.1. PROOF OF LEMMA 52: ISOTROPIC, NO AUGMENTATION

Under the stated setup, the covariance matrices in the formula evaluate to $\Sigma_o = \Sigma = \frac{1}{p} I_p$ and $\Sigma_* = I_p$. In this case, as $m = n$, $p \to \infty$ and $p/n \to \kappa$, we can compute the limit terms defined in (DO):

$$\begin{aligned}
\bar{\kappa}_* &= \lim_{p\to\infty} \frac{\|\beta^*\|}{\sqrt{p}} \;, & \bar{\kappa}_o &= \bar{\chi}_{11}^{\sigma,\tau} = 0 \;, & \bar{\chi}_{12}^{\sigma,\tau} &= \frac{\kappa}{\sigma_2 \tau_2 \lambda \kappa + 1} \;, & \bar{\chi}_{13}^{\sigma,\tau} &= \frac{1}{\sigma_2 \tau_2 \lambda \kappa + 1} \;, \\
\bar{\chi}_{21}^{\sigma,\tau} &= 0 \;, & \bar{\chi}_{22}^{\sigma,\tau} &= \frac{\kappa}{(\sigma_2 \tau_2 \lambda \kappa + 1)^2} \;, & \bar{\chi}_{23}^{\sigma,\tau} &= \frac{1}{(\sigma_2 \tau_2 \lambda \kappa + 1)^2} \;.
\end{aligned}$$

This implies

$$\bar{\chi}_1^{r,\theta,\sigma,\tau} = \frac{r_2^2 \kappa + \theta^2 \bar{\kappa}_*^2}{2(\lambda \kappa + \sigma_2^{-1} \tau_2^{-1})} \;, \qquad \bar{\chi}_2^{r,\theta,\sigma,\tau} = \frac{r_2^2 \kappa + \theta^2 \bar{\kappa}_*^2}{(\lambda \kappa + \sigma_2^{-1} \tau_2^{-1})^2} \;,$$

which are in particular independent of σ_1, τ_1 and r_1. Now recall that (EQs) read

$$\begin{cases} 0 \;=\; \theta\bar{\kappa}_*^2 - \frac{\alpha\bar{\kappa}_*^2}{\sigma_2\tau_2} - \frac{r_2\nu_2\bar{\kappa}_*}{k}\mathbb{E}\left[\bar{Z}_1\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta}\right] + r_2\nu_2\alpha\bar{\kappa}_*^2\;, \\ 0 \;=\; -\frac{1}{2\tau_1} - \partial_{\sigma_1}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau} + \frac{r_1\nu_1}{k}\,\mathbb{E}\Big[\eta^\intercal\left(I_k - \frac{1}{k}\mathbf{1}_{k\times k}\right)u_{\bar{Z},\varepsilon_1,\eta}\Big] + \frac{r_1\nu_1\sigma_1(k-1)}{k} + \frac{r_2\nu_2}{k}\mathbb{E}\Big[\eta^\intercal\frac{1}{k}\mathbf{1}_{k\times k}u_{\bar{Z},\varepsilon_1,\eta}\Big] \\ \qquad + \frac{r_2\nu_2\sigma_1}{k}\;, \\ 0 \;=\; -\frac{1}{2\tau_2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2^2\tau_2} - \partial_{\sigma_2}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau} + \frac{r_2\nu_2}{k}\mathbb{E}\Big[\bar{Z}_2\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta}\Big] + r_2\nu_2\sigma_2\;, \\ 0 \;=\; \frac{\sigma_1}{2\tau_1^2} - \partial_{\tau_1}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau}\;, \\ 0 \;=\; \frac{\sigma_2}{2\tau_2^2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2^2} - \partial_{\tau_2}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau}\;, \\ 0 \;=\; -\frac{r_1}{2\nu_1^2} + \frac{r_1}{2k}\,\mathbb{E}\Big[\Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1\eta)\Big\|^2\Big]\;, \\ 0 \;=\; -\frac{r_2}{2\nu_2^2} + \frac{1}{4r_2\nu_2^2} + \frac{r_2}{2k}\mathbb{E}\bigg[\bigg\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\bigg\|^2\bigg] \\ \qquad + \frac{1}{\nu_2 k}\mathbb{E}\bigg[\bar{Y}\Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2\nu_2}\bar{Y} - k\alpha\bar{\kappa}_*\bar{Z}_1\Big)\bigg]\;, \\ 0 \;=\; \frac{1}{2\nu_1} - \partial_{r_1}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau} + \frac{\nu_1}{2k}\,\mathbb{E}\Big[\Big\|(I_k - \frac{1}{k}\mathbf{1}_{k\times k})(u_{\bar{Z},\varepsilon_1,\eta} + \sigma_1\eta)\Big\|^2\Big]\;, \\ 0 \;=\; \frac{1}{2\nu_2} + \frac{1}{4r_2^2\nu_2} - \partial_{r_2}\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau} + \frac{\nu_2}{2k}\mathbb{E}\bigg[\bigg\|\frac{1}{k}\mathbf{1}_{k\times k}\Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y}\mathbf{1}_k - \alpha\bar{\kappa}_*\bar{Z}_1\mathbf{1}_k + \sigma_1\eta + \sigma_2\bar{Z}_2\mathbf{1}_k\Big)\bigg\|^2\bigg] \\ \qquad + \frac{1}{r_2 k}\mathbb{E}\bigg[\bar{Y}\Big(\mathbf{1}_k^\intercal u_{\bar{Z},\varepsilon_1,\eta} - \frac{k}{r_2\nu_2}\bar{Y} - k\alpha\bar{\kappa}_*\bar{Z}_1\Big)\bigg]\;, \\ 0 \;=\; \alpha\bar{\kappa}_*^2 - \partial_\theta\bar{\mathcal{X}}_1^{r,\theta,\sigma,\tau}\;. \end{cases}$$

By the 4th equation, $\sigma_1 = 0$. In this case, the defining optimization of $u_{\bar{Z},\varepsilon_1,\eta}$ is symmetric under permutation of $\tilde{u} \in \mathbb{R}^k$ and in particular $\frac{1}{k}\mathbf{1}_{k\times k}u_{\bar{Z},\varepsilon_1,\eta} = u_{\bar{Z},\varepsilon_1,\eta}$. This implies that $u_{\bar{Z},\varepsilon_1,\eta} = u_{\bar{Z},\varepsilon_1}\mathbf{1}_k$ where $u_{\bar{Z},\varepsilon_1}$ is the minimizer of the 1-d random optimization problem

$$\min_{\tilde{u}\in\mathbb{R}} \rho(\tilde{u}) + \frac{r_2\nu_2}{2}\big(\tilde{u} - \frac{1}{r_2\nu_2}\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2\big)^2\;. \tag{136}$$

Recall that $\mathrm{Prox}_{t\rho(\bullet)}(v) := \arg\min_{x\in\mathbb{R}} \frac{1}{2t}(v-x)^2 + \rho(x)$. This allows us to express

$$u_{\bar{Z},\varepsilon_1} \;= \mathrm{Prox}_{(r_2\nu_2)^{-1}\rho(\bullet)}\Big(\frac{1}{r_2\nu_2}\bar{Y} + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2\Big)\;.$$

Meanwhile, substituting $(I_k - \frac{1}{k}\mathbf{1}_{k\times k})u_{\bar{Z},\varepsilon_1,\eta} = 0$ into the 6th and 8th equations above yields $r_1 = 0$ and $\nu_1 \to \infty$. We can WLOG take $\nu_1 \to \infty$ such that $r_1\nu_1 \to 0$. By the 2nd equation we then obtain

$$\tau_1 \;=\; \frac{1}{2}\Big(\frac{r_2\nu_2}{k}\mathbb{E}[\eta^\intercal\mathbf{1}_k u_{\bar{Z},\varepsilon}]\Big)^{-1} \;\to\; \infty\;,$$

by noting that η is zero-mean and independent of $u_{\bar{Z},\varepsilon}$. This removes $(\sigma_1, r_1, \nu_1, \tau_1)$ from the equations. Substituting $u_{\bar{Z},\varepsilon_1,\eta} = \bar{u}_{\bar{Z},\varepsilon_1}\mathbf{1}_k$ and the derivatives of $\bar{\chi}_1^{r,\theta,\sigma,\tau}$, we obtain

$$\begin{cases} 0 \;=\; \theta\bar{\kappa}_*^2 - \frac{\alpha\bar{\kappa}_*^2}{\sigma_2\tau_2} - r_2\nu_2\bar{\kappa}_*\mathbb{E}\left[\bar{Z}_1 u_{\bar{Z},\varepsilon_1}\right] + r_2\nu_2\alpha\bar{\kappa}_*^2 \;, \\ 0 \;=\; -\frac{1}{2\tau_2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2^2\tau_2} - \frac{1}{\sigma_2^2\tau_2}\,\frac{r_2^2\kappa+\theta^2\bar{\kappa}_*^2}{2(\lambda\kappa+\sigma_2^{-1}\tau_2^{-1})^2} + r_2\nu_2\mathbb{E}[\bar{Z}_2 u_{\bar{Z},\varepsilon_1}] + r_2\nu_2\sigma_2 \;, \\ 0 \;=\; \frac{\sigma_2}{2\tau_2^2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2\tau_2^2} - \frac{1}{\sigma_2\tau_2^2}\,\frac{r_2^2\kappa+\theta^2\bar{\kappa}_*^2}{2(\lambda\kappa+\sigma_2^{-1}\tau_2^{-1})^2} \;, \\ 0 \;=\; -\frac{r_2}{2\nu_2^2} + \frac{1}{4r_2\nu_2^2} + \frac{r_2}{2}\mathbb{E}[(u_{\bar{Z},\varepsilon_1} - \frac{1}{r_2\nu_2}\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2] + \frac{1}{\nu_2}\mathbb{E}\Big[\bar{Y}\Big(u_{\bar{Z},\varepsilon_1} - \frac{1}{r_2\nu_2}\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \;, \\ 0 \;=\; \frac{1}{2\nu_2} + \frac{1}{4r_2^2\nu_2} - \frac{r_2\kappa}{\lambda\kappa+\sigma_2^{-1}\tau_2^{-1}} + \frac{\nu_2}{2}\mathbb{E}\Big[(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ \qquad + \frac{1}{r_2}\mathbb{E}\Big[\bar{Y}\Big(u_{\bar{Z},\varepsilon_1,\eta} - \frac{1}{r_2\nu_2}\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \;, \\ 0 \;=\; \alpha\bar{\kappa}_*^2 - \frac{\theta\bar{\kappa}_*^2}{\lambda\kappa+\sigma_2^{-1}\tau_2^{-1}} \;. \end{cases} \tag{137}$$

Now let $\gamma = \frac{1}{r_2\nu_2}$. Notice that the 4th and 5th equations above both involve

$$\begin{aligned} (\star) \;&:=\; \frac{1}{2}\mathbb{E}[(u_{\bar{Z},\varepsilon_1} - \gamma\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2] + \gamma\mathbb{E}\Big[\bar{Y}\Big(u_{\bar{Z},\varepsilon_1} - \gamma\bar{Y} - \alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \\ &\overset{(a)}{=}\; \mathbb{E}\Big[\frac{1-\bar{Y}}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ &\quad + \mathbb{E}\Big[\frac{\bar{Y}}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \gamma - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ &\quad + \gamma\mathbb{E}\Big[\bar{Y}\Big(\mathrm{Prox}_{\gamma\rho(\bullet)}(\gamma\bar{Y} + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \gamma - \alpha\bar{\kappa}_*\bar{Z}_1\Big)\Big] \\ &\overset{(b)}{=}\; \mathbb{E}\Big[\frac{\partial\rho(-\bar{\kappa}_*\bar{Z}_1)}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ &\quad + \mathbb{E}\Big[\frac{\partial\rho(\bar{\kappa}_*\bar{Z}_1)}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \gamma - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ &\quad - \gamma\mathbb{E}\Big[\partial\rho(\bar{\kappa}_*\bar{Z}_1)\Big(\alpha\bar{\kappa}_*\bar{Z}_1 + \gamma - \mathrm{Prox}_{\gamma\rho(\bullet)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)\Big)\Big] \\ &\overset{(c)}{=}\; \mathbb{E}\Big[\frac{\partial\rho(-\bar{\kappa}_*\bar{Z}_1)}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] \\ &\quad + \mathbb{E}\Big[\frac{\partial\rho(\bar{\kappa}_*\bar{Z}_1)}{2}(\mathrm{Prox}_{\gamma\rho(\bullet)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2) - \alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)^2\Big] - \frac{\gamma^2}{2} + \frac{\gamma^2}{4} \\ &\overset{(d)}{=}\; \mathbb{E}\Big[\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2 - \mathrm{Prox}_{\gamma\rho(\bullet)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))^2\Big] - \frac{\gamma^2}{4} \;. \end{aligned}$$

In (a) above, we have recalled that $\bar{Y} = \mathbb{I}_{\geq 0}\{\bar{\kappa}_o\bar{Z}_0 + \bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\} = \mathbb{I}_{\geq 0}\{\bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}$ is an indicator function; in (b) we have noted the equality of the conditional distributions $1 - \mathbb{I}_{\geq 0}\{\bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\,|\,\bar{Z}_1 \overset{d}{=} \mathbb{I}_{\geq 0}\{-\bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\,|\,\bar{Z}_1$ by the symmetry of ε_1 followed by $\sigma(\bullet) = \partial\rho(\bullet)$; in ($c$) we have expanded the square in the second term and noted that $\mathbb{E}[\partial\rho(\bar{\kappa}_*\bar{Z}_1)] = \mathbb{E}[(1 + e^{-\bar{\kappa}_*\bar{Z}_1})^{-1}] = \frac{1}{2}$ since $\bar{Z}_1$ is symmetric about zero; in (d), we have used in the second expectation that $\bar{Z}_1 \overset{d}{=} -\bar{Z}_1$, $\bar{Z}_2 \overset{d}{=} -\bar{Z}_2$ and that

$$\mathrm{Prox}_{\gamma\rho(\bullet)}\Big(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2\Big) \;=\; -\mathrm{Prox}_{\gamma\rho(\bullet)}\Big(-\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2\Big) \;,$$

where we have used $\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(x+\gamma) = -\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(-x)$ (see e.g. Lemma 3 of Salehi et al. (2019)). Substituting this into the last three lines of (137) gives

$$\begin{cases} \gamma^2 \;=\; \frac{2}{r_2^2}\mathbb{E}\Big[\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,(\bar{\kappa}_*\alpha\bar{Z}_1 + \sigma_2\bar{Z}_2 - \mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))^2\Big]\,, \\ \gamma \;=\; \frac{\kappa}{\lambda\kappa+\sigma_2^{-1}\tau_2^{-1}}\;, \\ \alpha \;=\; \frac{\theta}{\lambda\kappa+\sigma_2^{-1}\tau_2^{-1}}\;. \end{cases} \tag{138}$$

Meanwhile, the third line of (137) implies

$$\sigma_2^2 + \alpha^2\bar{\kappa}_*^2 \;=\; \frac{r_2^2\kappa + \theta^2\bar{\kappa}_*^2}{(\lambda\kappa + \sigma_2^{-1}\tau_2^{-1})^2} \;=\; \frac{r_2^2\kappa}{(\lambda\kappa + \sigma_2^{-1}\tau_2^{-1})^2} + \alpha^2\bar{\kappa}_*^2\;. \tag{139}$$

Combining the two calculations, we obtain

$$\theta \;=\; \frac{\alpha\kappa}{\gamma}\;, \qquad \tau_2 \;=\; \frac{\kappa^{-1}\gamma}{\sigma_2(1-\gamma\lambda)}\;, \qquad r_2 \;=\; \frac{\sigma_2\sqrt{\kappa}}{\gamma}\;, \tag{140}$$

which gives the first three desired equations. Substituting these back into the first line of (138) gives

$$\frac{\sigma^2\kappa}{2} \;=\; \mathbb{E}[\partial\rho(-\bar{\kappa}_*\bar{Z}_1)(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2 - \mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))^2]\,, \tag{141}$$

which is the fourth desired equation. The first and second equations of (137) are handled similarly as appendix C.3 of Salehi et al. (2019). We recall that $1 - \mathbb{I}_{\geq 0}\{\bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\,|\,\bar{Z}_1 \overset{d}{=} \mathbb{I}_{\geq 0}\{-\bar{\kappa}_*\bar{Z}_1 - \varepsilon_1\}\,|\,\bar{Z}_1$ and $\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(x+\gamma) = -\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(-x)$ again to compute

$$\begin{aligned} \mathbb{E}[\bar{Z}_1\, u_{\bar{Z},\varepsilon_1}] \;&=\; \mathbb{E}[\bar{Z}_1\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\gamma\bar{Y} + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] \\ &= \mathbb{E}[\bar{Z}_1\bar{Y}\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] + \mathbb{E}[\bar{Z}_1(1-\bar{Y})\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] \\ &= \mathbb{E}[\bar{Z}_1\,\partial\rho(\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\gamma + \alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] + \mathbb{E}[\bar{Z}_1\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] \\ &= \;-\,\mathbb{E}[\bar{Z}_1\,\partial\rho(\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\,-\,\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)] + \mathbb{E}[\bar{Z}_1\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] \\ &= 2\,\mathbb{E}[\bar{Z}_1\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] \\ &= \;-\,2\,\mathbb{E}[\bar{\kappa}_*\partial^2\rho(\,-\,\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2)] + \bar{\kappa}_*\alpha\,\mathbb{E}\bigg[\frac{\partial\rho(-\bar{\kappa}_*\bar{Z}_1)}{1 + \gamma\partial^2\rho(\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))}\bigg]\,, \end{aligned} \tag{142}$$

where the last line is exactly the same as (87)–(88) of Salehi et al. (2019) via Stein's lemma and by noting that $\bar{Z}_2 \overset{d}{=} -\bar{Z}_2$. Similarly

$$\mathbb{E}[\bar{Z}_2 u_{\bar{Z},\varepsilon_1}] = 2\,\mathbb{E}[\bar{Z}_2\,\partial\rho(-\bar{\kappa}_*\bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 - \sigma_2\bar{Z}_2)] = 2\sigma_2\,\mathbb{E}\bigg[\frac{\partial\rho(-\bar{\kappa}_*\bar{Z}_1)}{1 + \gamma\partial^2\rho(\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))}\bigg]\,, \tag{143}$$

where the last line is exactly the same as (83) of Salehi et al. (2019) via Stein's lemma. Substituting (143) into the second equation of (137) gives

$$0 \;=\; -\frac{1}{2\tau_2} + \frac{\alpha^2\bar{\kappa}_*^2}{2\sigma_2^2\tau_2} - \frac{1}{\sigma_2^2\tau_2}\,\frac{r_2^2\kappa + \theta^2\bar{\kappa}_*^2}{2(\lambda\kappa + \sigma_2^{-1}\tau_2^{-1})^2} + \frac{2\sigma_2}{\gamma}\,\mathbb{E}\bigg[\frac{\partial\rho(-\bar{\kappa}_*\bar{Z}_1)}{1 + \gamma\partial^2\rho(\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_*\bar{Z}_1 + \sigma_2\bar{Z}_2))}\bigg] + \frac{\sigma_2}{\gamma}\;.$$

Upon rearranging and a substitution of $\sigma_2^2 + \alpha^2 \bar{\kappa}_*^2 = \frac{r_2^2 \kappa + \theta^2 \bar{\kappa}_*^2}{(\lambda\kappa + \sigma_2^{-1}\tau_2^{-1})^2}$ from (139) and $\tau_2 \sigma_2 = \frac{\kappa^{-1}\gamma}{1-\gamma\lambda}$ from (140) , we obtain

$$1 - \frac{\gamma}{\tau_2 \sigma_2} \;=\; 1 - \kappa + \gamma\lambda\kappa \;=\; \mathbb{E}\Big[\frac{\partial\rho(-\bar{\kappa}_* \bar{Z}_1)}{1 + \gamma\partial^2\rho(\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_* \bar{Z}_1 + \sigma_2 \bar{Z}_2))} \Big] , \tag{144}$$

which gives the fifth desired equation. Substituting this into (142) implies

$$\mathbb{E}[\bar{Z}_1 u_{\bar{Z},\varepsilon_1}] \;=\; -\,2\,\mathbb{E}[\bar{\kappa}_* \partial^2 \rho(\,-\,\bar{\kappa}_* \bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_* \bar{Z}_1 + \sigma_2 \bar{Z}_2)] + \bar{\kappa}_* \alpha - \bar{\kappa}_* \alpha \,\frac{\gamma}{\tau_2 \sigma_2} \;,$$

and substituting this into the first equation of (137) gives

$$0 \;=\; \theta\bar{\kappa}_*^2 - \frac{\alpha\bar{\kappa}_*^2}{\sigma_2 \tau_2} - \frac{\bar{\kappa}_*}{\gamma}\Big(-\,2\,\mathbb{E}[\bar{\kappa}_* \partial^2 \rho(\,-\,\bar{\kappa}_* \bar{Z}_1)\,\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\alpha\bar{\kappa}_* \bar{Z}_1 + \sigma_2 \bar{Z}_2)] + \bar{\kappa}_* \alpha - \bar{\kappa}_* \alpha \,\frac{\gamma}{\tau_2 \sigma_2}\Big) + \frac{\alpha\bar{\kappa}_*^2}{\gamma} \;,$$

which simplifies to

$$-\frac{\gamma\theta}{2} \;= \mathbb{E}[\partial^2 \rho(-\bar{\kappa}_* \bar{Z}_1)\mathrm{Prox}_{\gamma\rho(\,\bullet\,)}(\bar{\kappa}_* \alpha \bar{Z}_1 + \sigma_2 \bar{Z}_2)] \;.$$

Replacing $\gamma\theta$ by $\alpha\kappa$ in view of (140) gives the last desired equation.

■

N.5.2. PROOF OF LEMMA 54: RANDOM PERMUTATIONS

Since $Z_1 \overset{d}{=} \phi_1(Z_1)$, Assumption 11(i) holds. Now note that by the total law of covariance followed by that ϕ_1 and ϕ_2 are i.i.d.,

$$\begin{aligned} \mathrm{Cov}\,[\phi_1(Z_1)\,,\,\phi_2(Z_1)] \;&= \mathrm{Cov}\,[\mathbb{E}[\phi_1(Z_1)\,|\,Z_1]\,,\,\mathbb{E}[\phi_2(Z_1)\,|\,Z_1]] \;+\; \mathbb{E}\,[\mathrm{Cov}[\phi_1(Z_1)\,,\,\phi_2(Z_1)\,|\,Z_1]] \\ &= \mathrm{Var}\,\mathbb{E}[\phi_1(Z_1)\,|\,Z_1]\;. \end{aligned}$$

Denote $\tilde{p}_t \;=\; \lceil r_{\mathrm{perm}} p_t \rceil$, the number of fixed entries of the l-th group to be permuted. We can WLOG suppose they are chosen as the first $\tilde{p}_t$ entries of the t-th group. Also write $Z^{(t)}_{t(\tilde{p}_t+1):tp_t} = (Z^{(t)}_{t(\tilde{p}_t+1)}, \ldots, Z^{(t)}_{tp_t})^\intercal$, the vector of un-permuted coordinates within the t-th group. Then we may compute

$$\begin{aligned} \Sigma_* \;&= (\Sigma^\dagger)^{1/2}\,\mathrm{Cov}\,[\phi_1(Z_1)\,,\,\phi_2(Z_1)]\,(\Sigma^\dagger)^{1/2} \\ &= (\Sigma^\dagger)^{1/2}\,\mathrm{Var}\,\mathbb{E}[\,\phi_1(Z_1)\,|\,Z_1\,]\,(\Sigma^\dagger)^{1/2} \\ &= (\Sigma^\dagger)^{1/2}\mathrm{Var}\begin{pmatrix} \frac{1}{\tilde{p}_1}\sum_{l\leq p_1} Z^{(1)}_{1l} \times \mathbf{1}_{\tilde{p}_1} \\ Z^{(1)}_{1(\tilde{p}_1+1):1p_1} \\ \vdots \\ \frac{1}{\tilde{p}_N}\sum_{l\leq \tilde{p}_N} Z^{(N)}_{Nl} \times \mathbf{1}_{\tilde{p}_N} \\ Z^{(N)}_{N(\tilde{p}_N+1):Np_N} \end{pmatrix}(\Sigma^\dagger)^{1/2} \\ &= (\Sigma^\dagger)^{1/2}\begin{pmatrix} \frac{1}{\tilde{p}_1}\,\mathrm{Var}[Z^{(1)}_{11}] \times \mathbf{1}_{\tilde{p}_1\times\tilde{p}_1} & & & & \\ & \mathrm{Var}[Z^{(1)}_{11}] \times I_{p-\tilde{p}_1} & & & \\ & & \ddots & & \\ & & & \frac{1}{\tilde{p}_N}\,\mathrm{Var}[Z^{(N)}_{11}] \times \mathbf{1}_{\tilde{p}_N\times\tilde{p}_N} & \\ & & & & \mathrm{Var}[Z^{(N)}_{11}] \times I_{p-\tilde{p}_N} \end{pmatrix}(\Sigma^\dagger)^{1/2} \;, \end{aligned}$$

whereas

$$\Sigma \;=\; \Sigma_o \;=\; \mathrm{Var}[Z_1] \;=\; \begin{pmatrix} \mathrm{Var}[Z_{11}^{(1)}]\times I_{p_1} & & \\ & \ddots & \\ & & \mathrm{Var}[Z_{11}^{(N)}]\times I_{p_N} \end{pmatrix},$$

and therefore

$$\Sigma_* \;=\; \begin{pmatrix} \frac{1}{\tilde{p}_1}\,\mathbf{1}_{\tilde{p}_1\times\tilde{p}_1}\,\mathbb{I}\{\mathrm{Var}[Z_{11}^{(1)}]>0\} & & & & \\ & I_{p-\tilde{p}_1}\,\mathbb{I}\{\mathrm{Var}[Z_{11}^{(1)}]>0\} & & & \\ & & \ddots & & \\ & & & \frac{1}{\tilde{p}_N}\,\mathbf{1}_{\tilde{p}_N\times\tilde{p}_N}\,\mathbb{I}\{\mathrm{Var}[Z_{11}^{(1)}]>0\} & \\ & & & & I_{p-\tilde{p}_N}\,\mathbb{I}\{\mathrm{Var}[Z_{11}^{(1)}]>0\} \end{pmatrix},$$

which satisfies $\Sigma_*^2 = \Sigma_*$. Thus Assumption 11(ii) holds. ■

N.5.3. PROOF OF LEMMA 55: RANDOM SIGN FLIPPING

Since $\Sigma = \Sigma_o = \frac{1}{p} I_p$, we can write

$$\Sigma_* \;= (\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_1(Z_1)\,,\,\phi_2(Z_1)]\,(\Sigma^\dagger)^{1/2} \;=\; \mathbb{E}[\phi_1]\,\mathbb{E}[\phi_2]$$

and

$$(\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_1(Z_1)\,,\,Z_1]\,(\Sigma_o^\dagger)^{1/2} \;=\; \mathbb{E}[\phi_1]\,.$$

WLOG we can suppose that the $\lceil r_{\text{flip}} p\rceil$ entries are chosen as the first $\lceil r_{\text{flip}} p\rceil$ entries. Then each $\phi_{ij} = \mathrm{diag}\{\mathrm{Rad}_{ij1}, \ldots, \mathrm{Rad}_{ij\lceil r_{\text{flip}} p\rceil}, 1, \ldots, 1\}$, where Rad_{ijl}'s are i.i.d. Rademacher random variables. Therefore $\mathbb{E}[\phi_1] = \mathrm{diag}\{0, \ldots, 0, 1, \ldots, 1\}$, where there are $\lceil r_{\text{flip}} p\rceil$ zeros, and in particular $\mathbb{E}[\phi_1] = \mathbb{E}[\phi_1]\mathbb{E}[\phi_1] = \mathbb{E}[\phi_1]\mathbb{E}[\phi_2^\top]$. This verifies both Assumption 11(i) and (ii). ■

N.5.4. PROOF OF LEMMA 56: RANDOM CROPPING

In the random cropping setup, $\Sigma_o = \Sigma_{\text{new}} = \frac{1}{p} I_p$. Also note that each ϕ_i is a random projection matrix and independent of Z_i. Then by the total law of covariance,

$$\begin{aligned} \Sigma \;=\; \mathrm{Var}[\phi_1(Z_1)] \;&= \mathbb{E}\mathrm{Var}[\phi_1(Z_1)\,|\,\phi_1] + \mathrm{Var}\mathbb{E}[\phi_1(Z_1)\,|\,\phi_1] \\ &= \mathbb{E}[\phi_1 \mathrm{Var}[Z_1]\phi_1] + 0 \;=\; \frac{1}{p}\mathbb{E}[\phi_1] \;=\; \frac{1}{p}\,\frac{p-\lceil r_{\text{crop}} p\rceil}{p}\,I_p\,. \end{aligned}$$

This implies

$$\begin{aligned} \Sigma_* \;&= (\Sigma^\dagger)^{1/2}\,\mathrm{Cov}[\phi_1(Z_1), \phi_2(Z_1)]\,(\Sigma^\dagger)^{1/2} \\ &= \Big(\frac{1}{p}\,\frac{p-\lceil r_{\text{crop}} p\rceil}{p}\Big)^{-1}\,\mathbb{E}[\phi_1]\,\mathrm{Var}[Z_1]\,\mathbb{E}[\phi_2] \;=\; \frac{p-\lceil r_{\text{crop}} p\rceil}{p}\,I_p\,, \end{aligned}$$

and

$$(\Sigma^\dagger)^{1/2}\mathrm{Cov}[\phi_1(Z_1), Z_1](\Sigma_o^\dagger)^{1/2} \;= (\Sigma^\dagger)^{1/2}\mathbb{E}[\phi_1]\mathrm{Var}[Z_1](\Sigma_o^\dagger)^{1/2} \;=\; I_p\,.$$

Therefore the desired statements hold with $a_1 = a_2 = \frac{p-\lceil r_{\text{crop}} p\rceil}{p}$. ■