A Novel Convex Gaussian Min Max Theorem for Repeated Features

Citation for the original published paper (version of record):
Bosch, D., Panahi, A. (2025). A Novel Convex Gaussian Min Max Theorem for Repeated Features. Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence, UAI 2022, 258: 3673-3681

A Novel Convex Gaussian Min Max Theorem for Repeated Features

David Bosch **Ashkan Panahi**
Chalmers University of Technology and University of Gothenburg

Abstract

The Convex Gaussian Min-Max Theorem (CGMT) is a powerful method for the study of min-max optimization problems over bilinear Gaussian forms. It provides an alternative optimization problem whose statistical properties are tied to that of the target problem. We prove a generalization of the CGMT to a family of problems in machine learning (ML) with correlated entries in the data matrix. This family includes various familiar examples of problems with shared weights or repeated features. We make use of our theorem to obtain asymptotically exact learning curves for regression with vector-valued labels, complex variables, and convolution.

1 INTRODUCTION

Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be a random matrix and $\mathcal{S}_1 \subseteq \mathbb{R}^m, \mathcal{S}_2 \subseteq \mathbb{R}^n$ two sets. Then, for a given function $\psi : \mathcal{S}_1 \times \mathcal{S}_2 \to \mathbb{R}$, consider the following min-max optimization problem

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_1} \max_{\mathbf{z}\in\mathcal{S}_2} \mathbf{z}^T\mathbf{X}\boldsymbol{\theta} + \psi(\boldsymbol{\theta}, \mathbf{z}). \tag{1}$$

A multitude of problems in statistics, machine learning and AI can be expressed by (1), and its analysis for large dimensions m, n has recently received much attention. While studying (1) in complete generality is extremely difficult, in many common cases, especially for convex-concave programs, precise descriptions of this problem are readily available. Many of these results rely on the Convex Gaussian Min-Max Theorem (CGMT) (Thrampoulidis et al., 2014), which provides a fairly straightforward way to obtain computable expressions for the asymptotic properties of the solution of (1) and provable concentration bounds on them. However, the CGMT is limited to random matrices with i.i.d. entries. As a result, it is inapplicable to problems with shared or repeated features that occur even in simple cases such as regression with vector-valued labels or convolutional models. In this work, we resolve this limitation by extending the CGMT to a wide range of setups with statistical correlation among the elements of $\mathbf{G}$, including sharing of the features. Our main contributions are summarized below:

1. We resolve the limitation of the CGMT requiring i.i.d. elements of the data matrix by proving a more general theorem, of which the CGMT is a special case. In contrast to the conventional proof of the CGMT by Gordon's comparison theorem, we make use of an approach pioneered by Stojnic (2016a,b), who has previously demonstrated that both Slepian's lemma and Gordon's comparison theorem can be expressed and derived through a single framework. We extend those results to the case of weight sharing, proving both a "Slepian" style expression, concerning $\max-\max$ optimization in the form of (1), and a "Gordon" style expression involving $\min-\max$ optimization.
2. We make use of our theorem to analyze the problems which cannot be analyzed by the classical CGMT. These problems include regression with vector labels, regression with convolutions, and regression over complex variables. We verify our claims experimentally, showing a match between the primary and CGMT alternative optimizations, and show that our theory can predict specific phenomena, such as double descent for vector labeled regression and complex regression.

1.1 Relevance to Machine Learning:

We consider the analysis of linear models as a prominent use case of the CGMT in machine learning. Consider such a model parameterized by a real vector $\boldsymbol{\theta} \in \mathbb{R}^d$, a loss function ℓ, regularization function R, and a dataset $\{(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^n$. A linear model parameterized by $\boldsymbol{\theta}$, which is a predictor of y, is commonly obtained by solving the following empirical risk

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

minimization problem:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x}_i^T\boldsymbol{\theta}, y_i) + R(\boldsymbol{\theta}). \tag{2}$$

It is well-known that (2) can be expressed in the form of (1), by means of a Legendre transform, ℓ^*, of the loss function with respect to the first element, (Bauschke et al., 2011):

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d}\max_{\mathbf{z}\in\mathbb{R}^n} \frac{1}{n}\mathbf{z}^T\mathbf{X}\boldsymbol{\theta} - \frac{1}{n}\sum_{i=1}^{n}\ell^*(z_i, y_i) + R(\boldsymbol{\theta}), \tag{3}$$

where the matrix $\mathbf{X}\in\mathbb{R}^{n\times d}$ is the collection of the samples $\mathbf{x}_i$. In (3), the transformed loss and regularization function amount to ψ in (1).

1.1.1 Application of CGMT

If $\mathbf{X}$ has i.i.d. standard Gaussian elements, then under mild assumptions on the loss and regularizer this problem is amenable to the CGMT, offering the following alternative optimization to (1):

$$\min_{\boldsymbol{\theta}\in\mathcal{S}_1}\max_{\mathbf{z}\in\mathcal{S}_2} \|\boldsymbol{\theta}\|\,\mathbf{g}^T\mathbf{z} + \|\mathbf{z}\|\,\mathbf{h}^T\boldsymbol{\theta} + \psi(\boldsymbol{\theta}, \mathbf{z}), \tag{4}$$

where $\mathbf{g}\in\mathbb{R}^n, \mathbf{h}\in\mathbb{R}^m$ are i.i.d. standard Gaussian vectors. In simple words, the CGMT establishes that the solutions of (4) and (1) are identical in a wide range of asymptotic statistical properties. Calculating these properties for (4) is significantly simpler than the original problem in (1). If $\mathbf{X}$ is i.i.d., but not Gaussian, there are universality arguments, which under certain assumptions on the loss function, regularizer, and the p.d.f. of $\mathbf{X}$, demonstrate that the same expressions as in the Gaussian case, hold valid for the non-Gaussian case (Panahi and Hassibi, 2017; Hu and Lu, 2022; Bosch et al., 2023; Han and Shen, 2023). This makes CGMT a powerful tool of analysis, even for non-Gaussian data.

1.1.2 Limitation of CGMT

For the analysis of more complex problems, the central limitation of the CGMT is the requirement that the elements of $\mathbf{G}$ are i.i.d. For example, consider the case of linear regression with vector-valued labels of dimension k:

$$\min_{\boldsymbol{\Theta}\in\mathbb{R}^{m\times k}} \frac{1}{2n}\|\mathbf{X}\boldsymbol{\Theta} - \mathbf{Y}\|_F^2 + R(\boldsymbol{\Theta}), \tag{5}$$

where we assume the data matrix $\mathbf{X}\in\mathbb{R}^{n\times m}$ has i.i.d. Gaussian entries, $\mathbf{Y}\in\mathbb{R}^{n\times k}$ is the collection of vector-valued labels $\mathbf{y}_i\in\mathbb{R}^k$, and $\|\cdot\|_F$ denotes the Frobenius norm. Invoking the Legendre transform again, this problem can be expressed in the form of (1) as follows:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{mk}}\max_{\mathbf{z}\in\mathbb{R}^{nk}} \frac{1}{n}\mathbf{z}^T\left(\mathbf{X}\otimes\mathbf{I}_k\right)\boldsymbol{\theta} - \frac{1}{n}\mathbf{z}^T\mathbf{y} - \frac{1}{2n}\|\mathbf{z}\|_2^2 + R(\mathrm{vec}^{-1}(\boldsymbol{\theta})), \tag{6}$$

where $\mathbf{I}_k$ is the identity matrix of size k, $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\Theta})$, $\mathbf{z} = \mathrm{vec}(\mathbf{Z})$ and $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, where vec denotes the vectorization operation, and $\otimes$ denotes the kronecker product. This problem is of the form of (1), but cannot be analyzed by means of currently existing comparison theorems as the matrix $\mathbf{X}\otimes\mathbf{I}_k$, is no longer i.i.d. and instead repeats elements of $\mathbf{X}$ many times over. In the rest of this paper, we develop a generalized result that can address this problem and many similar ones.

2 RELATED WORKS

The approach of Gaussian comparison theorems stems from the seminal work of Slepian (1962), which relates the maximum value of a pair of Gaussian processes with specific relations between their covariance functions. Slepian also introduced the following instance of the Gaussian pairs to achieve sharp bounds on the operator norm of Gaussian matrices:

$$\mathbf{x}^T\mathbf{G}\mathbf{y} + \gamma\|\mathbf{x}\|\|\mathbf{y}\|, \qquad \|\mathbf{x}\|\,\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|\,\mathbf{h}^T\mathbf{x}, \tag{7}$$

where $\mathbf{G}\in\mathbb{R}^{n\times m}, \mathbf{g}\in\mathbb{R}^m, \mathbf{h}\in\mathbb{R}^n, \gamma\in\mathbb{R}$ are i.i.d. Gaussian. Gordon (1985, 1988) extended Slepian's comparison theorem, and showed that the same pair of primary and alternative Gaussian processes are also linked in terms of their $\min_{\mathbf{x}}\max_{\mathbf{y}}$ values. Applications of this theorem include, (Rudelson and Vershynin, 2006; Stojnic, 2013a; Oymak et al., 2013; Thrampoulidis et al., 2014, 2015). Subsequently, Thrampoulidis et al. (2014), based on observations by earlier work (Stojnic, 2013b; Amelunxen et al., 2013), demonstrated that under the additional assumptions of convexity, the bounds coincide with exact asymptotic optimal values in the Gordon and Slepian results. This so-called convex Gaussian min-max theorem (CGMT) has been used to study a wide range of problems in signal processing and machine learning, including but not limited to (Akhtiamov et al., 2023; Aolaritei et al., 2023; Javanmard and Soltanolkotabi, 2022; Mignacco et al., 2020; Montanari et al., 2019; Salehi et al., 2019; Thrampoulidis et al., 2018; Zhou et al., 2024; Loureiro et al., 2021; Bosch et al., 2022).

Several studies have attempted to address the central limitation of the CGMT, namely the i.i.d. requirement on the matrix $\mathbf{G}$. Thrampoulidis et al. (2020) studied the problem of multiclass regression by considering a set of pairwise CGMTs between each pair of classes. Dhifallah and Lu (2021) considers an extension to the case of a sum of bilinear Gaussian forms;

$\sum_{k=1}^{K} \mathbf{x}_k^T \mathbf{G}_k \mathbf{y}_k$, where each $\mathbf{G}_k$ is i.i.d. Gaussian and independent. Recently, Akhtiamov et al. (2024) considered a further generalization to the case where the bilinear Gaussian form is given by; $\sum_{k=1}^{K} \mathbf{x}_k^T \mathbf{G}_k \mathbf{\Sigma}_k^{1/2} \mathbf{y}$; where $\mathbf{y}$ is shared between all K terms, and each $\mathbf{G}_k$ is i.i.d. Gaussian and independent and $\mathbf{\Sigma}_k^{1/2}$ are positive semi-definite covariance matrices. Our approach is a more general extension to a wider class of weight-sharing setups.

Our approach is based on an alternative proof to the Gaussian min-max theorem, as well as a "Slepian" variant concerning a max-max optimization theorem, pioneered by Stojnic (2016a,b). This study proves both of these results through a single argument, called random duality theory (RDT). More recently, Stojnic (2023a,b,c) has extended these results to optimization problems with random linear constraints. These results however do still require that the random constraints are i.i.d. Gaussian, and cannot be used to consider cases of weight sharing, such as those discussed in this work.

3 MAIN RESULTS

3.1 Notation

For any natural number $K \in \mathbb{N}$ we will use $[K]$ to denote the set $\{1, 2, \dots, K\}$ of natural numbers up to and including K. We will denote the Kronecker delta by $\delta_{a,b}$, where $a, b \in \mathbb{N}$, defined to be 1 if $a = b$ and 0 otherwise. We will denote vectors using boldface lower case letters, such as $\mathbf{a}, \mathbf{b}, \mathbf{c}$, and matrices by boldface upper case letters, ie. $\mathbf{A}, \mathbf{B}, \mathbf{C}$. When considering a particular element of a vector or matrix, we will use a non-bold font and subscript the object, for example, $A_{a,b}$ means the (a, b) element of the matrix $\mathbf{A}$. By contrast, boldfaced letters with subscripts will denote an element of a set of matrices or vectors, for example, $\mathbf{A}_i$ means the ith matrix of a set of matrices, e.g. $\{\mathbf{A}_i\}_{i=1}^{n}$. When considering an element of a set of matrices, we use the notation $(\mathbf{A}_i)_{a,b}$ to refer to the (a, b)th element, of the ith matrix of a set of matrices.

For any natural number $K \in \mathbb{N}$, we denote by $\mathcal{S}_K^+ \subset \mathbb{R}^{K \times K}$ the cone of symmetric Positive Semi-Definite (PSD) matrices of size $K \times K$. For a PSD matrix $\mathbf{A}$, we will denote by $\mathbf{A}^{1/2}$ the unique PSD matrix square root. For any natural number $K \in \mathbb{N}$, the matrices $\mathbf{I}_K$ and $\mathbf{0}_K$ denote the identity matrix and all zeros matrices of size $K \times K$, respectively. For a PSD matrix $\mathbf{A}$, vector $\mathbf{b}$ and function f, we denote by

$$\mathcal{M}_{\mathbf{A}\ f}(\mathbf{b}) = \min_{\mathbf{x}} f(\mathbf{x}) + \frac{1}{2}(\mathbf{x} - \mathbf{b})^T \mathbf{A}(\mathbf{x} - \mathbf{b}), \qquad (8)$$

the matrix version of the Moreau envelope. Correspondingly, $\text{prox}_{\mathbf{A}\ f}(\mathbf{b})$ is the proximal operator, referring to the optimal point of the optimization in equation (8).

3.2 Extension to the Gaussian Min-Max Theorem

Our result concerns a particular pattern of statistical correlation in $\mathbf{G}$ that we refer to as *Gaussian Matrix Sum (GMS)*. We first formalize this idea:

Definition 1 (Gaussian Matrix Sum (GMS))**.** Let $\tilde{\mathbf{G}} \in \mathbb{R}^{\tilde{n} \times \tilde{m}}$ be an i.i.d. standard Gaussian matrix, let $K \in \mathbb{N}$ be a positive integer, and let $\mathbf{A}_k \in \mathbb{R}^{\tilde{n} \times n}, \mathbf{B}_k \in \mathbb{R}^{\tilde{m} \times m}$ for $k \in [K]$, be two sets of deterministic matrices. Then $\mathbf{G}$ is a *Gaussian Matrix Sum (GMS)* of order K and with components $\{\mathbf{A}_k, \mathbf{B}_k\}_{k=1}^{K}$, if

$$\mathbf{G} = \sum_{k=1}^{K} \mathbf{A}_k^T \tilde{\mathbf{G}} \mathbf{B}_k. \qquad (9)$$

In a GMS, the components $\mathbf{A}_k, \mathbf{B}_k$ express the manner in which the weights of the Gaussian matrix $\tilde{\mathbf{G}}$ are shared in the total matrix $\mathbf{G}$. In section 5, we show how common cases, such as multiclass regression can be expressed as GMSs.

Given this definition, we state our main result for this section, which consists of two claims: first, an extension of the Gaussian Min-Max Theorem (GMT) (Gordon, 1985; Thrampoulidis et al., 2014) and second, an extension of Slepian's Lemma (Slepian, 1962). For completeness, we state the original comparison theorems in appendix A.

We take a Gaussian Matrix Sum $\mathbf{G} \in \mathbb{R}^{n \times m}$ of order K and components $\{\mathbf{A}_k, \mathbf{B}_k\}_{k=1}^{K}$, as defined in definition 1. Further, we consider two compact sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$, and a continuous function $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. We define the following *primary* objective function

$$\mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \boldsymbol{\gamma}, \psi) = \mathbf{x}^T \mathbf{G} \mathbf{y} + \text{Tr}[\mathbf{P}^{1/2} \boldsymbol{\gamma} \mathbf{Q}^{1/2}] + \psi(\mathbf{x}, \mathbf{y}), \qquad (10)$$

where $\mathbf{x} \in \mathcal{X}, \mathbf{y} \in \mathcal{Y}$, $\boldsymbol{\gamma} \in \mathbb{R}^{K \times K}$ is an i.i.d. standard Gaussian matrix independent of $\mathbf{G}$, and $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{K \times K}$ are defined element-wise as follows:

$$P_{k,k'} = \mathbf{x}^T \mathbf{A}_k^T \mathbf{A}_{k'} \mathbf{x}, \qquad Q_{k,k'} = \mathbf{y}^T \mathbf{B}_k^T \mathbf{B}_{k'} \mathbf{y}. \qquad (11)$$

for $k, k' \in [K]$. We note that $\mathbf{P}$ and $\mathbf{Q}$ are positive semi-definite (PSD) matrices, and hence $\mathbf{P}^{1/2}, \mathbf{Q}^{1/2} \in \mathbb{R}^{K \times K}$ are the unique PSD square roots of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Next, we define the following *alternative* objective function

$$\mathcal{H}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi) = \sum_{k=1}^{K} \mathbf{f}_k^T \mathbf{B}_k \mathbf{y} + \mathbf{h}_k^T \mathbf{A}_k \mathbf{x} + \psi(\mathbf{x}, \mathbf{y}), \qquad (12)$$

where $\mathbf{F} \in \mathbb{R}^{\tilde{m}\times K}, \mathbf{H} \in \mathbb{R}^{\tilde{n}\times K}$ are Gaussian matrices with columns $\mathbf{f}_k \in \mathbb{R}^{\tilde{m}}, \mathbf{h}_k \in \mathbb{R}^{\tilde{n}}$ for $k \in [K]$, respectively. $\mathbf{F}$ and $\mathbf{H}$ have zero mean, and are defined by:

$$\mathbf{F} = \tilde{\mathbf{F}}\mathbf{P}^{1/2}, \qquad \mathbf{H} = \tilde{\mathbf{H}}\mathbf{Q}^{1/2}, \tag{13}$$

where $\tilde{\mathbf{F}} \in \mathbb{R}^{\tilde{m}\times K}, \tilde{\mathbf{H}} \in \mathbb{R}^{\tilde{n}\times K}$ have i.i.d. standard Gaussian entries.

Having defined the primary and alternative objectives, we now express the following extension to the Gaussian Comparison Theorems.

Theorem 1 (Generalization to Slepian comparison and Gaussian Min Max Theorem)**.** *Let $\mathcal{X} \subset \mathbb{R}^n, \mathcal{Y} \subset \mathbb{R}^m$ be compact sets, and consider the primary $\mathcal{H}_\mathcal{P}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \boldsymbol{\gamma}, \psi)$ and alternative $\mathcal{H}_\mathcal{A}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi)$ objectives as defined in equations* (10) *and* (12) *respectively. Furthermore, assume that $\psi(\mathbf{x}, \mathbf{y})$ is continuous and* strictly concave *in $\mathbf{y}$ and $\mathbf{x}$. Then:*

$$\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_\mathcal{P}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \boldsymbol{\gamma}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_\mathcal{A}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi).$$

Alternatively, assume that $\psi(\mathbf{x}, \mathbf{y})$ is continuous, strictly concave *in $\mathbf{y}$ and jointly* strictly convex *in $\mathbf{x}$. Then:*

$$\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_\mathcal{P}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \boldsymbol{\gamma}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_\mathcal{A}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi).$$

Furthermore, let $(\hat{\mathbf{x}}_\mathcal{P}, \hat{\mathbf{y}}_\mathcal{P})$ and $(\hat{\mathbf{x}}_\mathcal{A}, \hat{\mathbf{y}}_\mathcal{A})$ denote the optimal points of $\mathcal{H}_\mathcal{P}$ and $\mathcal{H}_\mathcal{A}$ respectively, for either the max-max or min-max optimization. Then,

$$Var_{\mathbf{G},\boldsymbol{\gamma}}[\mathcal{H}_\mathcal{P}(\hat{\mathbf{x}}_\mathcal{P}, \hat{\mathbf{y}}_\mathcal{P}, \mathbf{G}, \boldsymbol{\gamma}, \psi)] = Var_{\mathbf{F},\mathbf{H}}[\mathcal{H}_\mathcal{A}(\hat{\mathbf{x}}_\mathcal{A}, \hat{\mathbf{y}}_\mathcal{A}, \mathbf{F}, \mathbf{H}, \psi)].$$

Remark 1. We note that the primary and alternative objectives of the GMT are a special case of this theorem, where $K = 1$, and $\mathbf{A} = \mathbf{I}_n, \mathbf{B} = \mathbf{I}_m$. In this case $\mathbf{P} \in \mathbb{R}^{1\times 1} = \|\mathbf{x}\|^2$ and $\mathbf{Q} \in \mathbb{R}^{1\times 1} = \|\mathbf{y}\|^2$, which results in the familiar pair of primary and alternative objectives given by:

$$\mathcal{H}_\mathcal{P} = \mathbf{x}^T\mathbf{G}\mathbf{y} + \|\mathbf{x}\|\,\|\mathbf{y}\|\,\gamma + \psi(\mathbf{x}, \mathbf{y}),$$
$$\mathcal{H}_\mathcal{A} = \|\mathbf{x}\|\,\mathbf{f}^T\mathbf{y} + \|\mathbf{y}\|\,\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x}, \mathbf{y}).$$

3.2.1 Proof Sketch

The two statements of theorem 1 are proved together in a single master theorem. To prove the equivalence of the expected values, we consider an interpolation between $\mathcal{H}_\mathcal{P}$ and $\mathcal{H}_\mathcal{A}$:

$$\mathcal{H}_t = \sqrt{1-t}\mathcal{H}_\mathcal{P} + \sqrt{t}\mathcal{H}_\mathcal{A}, \tag{14}$$

where $t \in [0, 1]$, such that at $t = 0$ we have that $\mathcal{H}_0 = \mathcal{H}_\mathcal{P}$ and at $t = 1$ we have $\mathcal{H}_1 = \mathcal{H}_\mathcal{A}$. We then re-express the max − max or min − max optimization over $\mathcal{H}_t$ as the low-temperature limit of a soft-max (Boltzmann distribution) over the sets $\mathcal{X}, \mathcal{Y}$:

$$\xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, t) = \mathbb{E}_{\tilde{\mathbf{G}},\boldsymbol{\gamma},\mathbf{F},\mathbf{H}} \tfrac{1}{\beta|s|} \log\left(\int_\mathcal{X} d\mathbf{x} \left(\int_\mathcal{Y} d\mathbf{y}\; e^{\beta\mathcal{H}_t(\mathbf{x},\mathbf{y},\psi)}\right)^s\right) \tag{15}$$

Here $\beta > 0$ corresponds to the "inverse temperature" and $s \in \{-1, 1\}$ is a parameter. In the limit of $\beta \to \infty$, we observe that

$$\lim_{\beta\to\infty} \xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, t) = \mathbb{E}_{\mathbf{G},\boldsymbol{\gamma},\mathbf{F},\mathbf{H}} \max_{\mathbf{x}\in\mathcal{X}} s \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi). \tag{16}$$

Then, we see that taking $s = 1$ results in a max − max optimization and $s = -1$ to a min − max optimization. The proof proceeds by considering the $t-$derivative of the function ξ, where eventually we show (see supplement section B.3) that

$$\lim_{\beta\to\infty} \frac{d\xi}{dt} = 0. \tag{17}$$

Then, we conclude by the dominated convergence theorem that:

$$\lim_{\beta\to\infty} \xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, 0) = \lim_{\beta\to\infty} \xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, 1), \tag{18}$$

which implies the desired results. For the variance, we consider the expected value of the observable $\mathcal{H}_t^2$ with respect to both the Boltzmann distribution and the Gaussian terms $\{\mathbf{G}, \boldsymbol{\gamma}, \mathbf{F}, \mathbf{H}\}$. We once again take the derivative of this observable with respect to t and show that in the large β limit the derivative goes to zero. The full proof is found in appendix B.

3.3 Extension to the Convex Gaussian Min-Max Theorem

We note that the previous theorem requires the existence of the $\boldsymbol{\gamma}$ term in the primary optimization. As discussed above, in the majority of the optimization problems of interest, this term is not present. In the following theorem below, we show that the results of the theorem continue to hold even if this term is removed.

Theorem 2 (Generalization of the Convex Gaussian Min-Max Theorem)**.** *Assume the setup of Theorem 1, and consider instead the following primary objective function:*

$$\mathcal{H}_\mathcal{R}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \psi) = \mathbf{x}^T\mathbf{G}\mathbf{y} + \psi(\mathbf{x}, \mathbf{y}). \tag{19}$$

Then assume that $\psi(\mathbf{x}, \mathbf{y})$ *is continuous and* strictly concave *in* $\mathbf{x}$ *and* $\mathbf{y}$*. Then,*

$$\mathbb{E}_{\mathbf{G}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi).$$

Assume instead that $\psi(\mathbf{x}, \mathbf{y})$ *is continuous,* strictly convex *in* $\mathbf{x}$ *and* strictly concave *in* $\mathbf{y}$*. Then,*

$$\mathbb{E}_{\mathbf{G}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi).$$

Furthermore, let $(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}})$ *and* $(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}})$ *denote the optimal points of* $\mathcal{H}_{\mathcal{R}}$ *and* $\mathcal{H}_{\mathcal{A}}$ *respectively, for either the max-max or min-max optimization. Then,*

$$Var_{\mathbf{G}}[\mathcal{H}_{\mathcal{R}}(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}}, \mathbf{G}, \psi)] \leq Var_{\mathbf{F},\mathbf{H}}[\mathcal{H}_{\mathcal{A}}(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}}, \mathbf{F}, \mathbf{H}, \psi)].$$

A proof can be found in the supplement section C.

We note that the results of theorems 1 and 2 are not asymptotic. The expected values match, regardless of the dimension of the problem. This is in contrast to the CGMT which provides bounds on the relationship between the optimal values, and tightness of the bounds requires the concentration of optimal values. Our theorem, of which the objectives considered by the CGMT are a special case, shows that the expected values of the objectives will match even non-asymptotically if strict convexity is assumed.

While the theorems above show that the expected values of the objectives match at the optimal points, it does not prove any relationship between the distribution of the optimal points, themselves. In the following theorem, we show that many functions of the optimal solutions of $\mathcal{H}_{\mathcal{P}}, \mathcal{H}_{\mathcal{A}}$, and $\mathcal{H}_{\mathcal{R}}$ will also match in expectation. This allows for more general properties of the optimization problems in question to be analyzed by proxy. For example, in the case of empirical risk minimization, the expected generalization error can be predicted by means of the optimal solutions to the alternative optimization problem, instead of those of the primary.

Theorem 3 (Functions of Solutions)**.** *Assume the setup of Theorem 1, and let* $(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}})$ *and* $(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}})$*, denote the optimal points of* $\mathcal{H}_{\mathcal{P}}$ *and* $\mathcal{H}_{\mathcal{A}}$ *respectively. Furthermore, assume that* $\mathcal{H}_{\mathcal{P}}$ *and* $\mathcal{H}_{\mathcal{A}}$ *have finite values at their optimal points, and that* ψ *is a continuous function.*

Let $\phi(\mathbf{x}, \mathbf{y}) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *be any bounded continuous function, ie.* $\|\phi\|_{\infty} < \infty$*. Then,*

$$\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}[\phi(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}})] = \mathbb{E}_{\mathbf{F},\mathbf{H}}[\phi(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}})]. \tag{20}$$

Furthermore, assuming the conditions of Theorem 2, and denoting by $(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}})$ *the optimal point of* $\mathcal{H}_{\mathcal{R}}$*, and assuming that* $\mathcal{H}_{\mathcal{R}}$ *is finite at the optimal point, we can similarly find that:*

$$\mathbb{E}_{\mathbf{G}}[\phi(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}})] = \mathbb{E}_{\mathbf{F},\mathbf{H}}[\phi(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}})]. \tag{21}$$

A proof is given in the supplement section D.

4 SPECIAL CASES

In this section we consider two specific cases for the shape of the GMS $\mathbf{G}$ and show that in these cases the alternative optimization form, as described in (12), can be re-expressed into more convenient and intuitive forms. We first consider the case in which the quadratic form $\mathbf{x}^T\mathbf{G}\mathbf{y}$ is a re-expression of a trace. This is, for example, the case in regression with vector-valued labels, discussed in more detail below.

Corollary 1 (Trace Form)**.** Consider the compact sets $\mathcal{X} \subset \mathbb{R}^{n\times k}$ and $\mathcal{Y} \subset \mathbb{R}^{m\times k}$, and the following primary objective function:

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{\mathbf{Y}\in\mathcal{Y}} \mathrm{Tr}[\mathbf{X}^T\mathbf{G}\mathbf{Y}] + \psi(\mathbf{X}, \mathbf{Y}). \tag{22}$$

Then the corresponding alternative optimization is given by:

$$\min_{\mathbf{X}\in\mathcal{X}} \max_{\mathbf{Y}\in\mathcal{Y}} \mathrm{Tr}[(\mathbf{X}^T\mathbf{X})^{1/2}\mathbf{F}\mathbf{Y}] + \mathrm{Tr}[(\mathbf{Y}^T\mathbf{Y})^{1/2}\mathbf{H}\mathbf{X}] + \psi(\mathbf{X}, \mathbf{Y}), \tag{23}$$

where $\mathbf{F} \in \mathbb{R}^{k\times m}$ and $\mathbf{H} \in \mathbb{R}^{k\times n}$ are i.i.d. standard Gaussian matrices.

This is a direct extension of the ordinary CGMT to the case of vector-valued labels, and we can see that if $k = 1$, this alternative expression exactly matches the alternative of the classical CGMT. A proof of this fact is given in the supplement section E.1.

A second special case that we consider is the case where $\mathbf{G}$ takes an alternating form. This form shows up in the analysis of regression over complex variables, which we also discuss in more detail below.

Corollary 2 (Complex Form)**.** Consider two compact sets $\mathcal{X} \subset \mathbb{R}^{2n}$ and $\mathcal{Y} \subset \mathbb{R}^{2m}$, let $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{n\times m}$ be two i.i.d. Standard Gaussian matrices. Then consider the following primary optimization:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{G} & -\mathbf{H}\\ \mathbf{H} & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} + \psi(\mathbf{x}, \mathbf{y}), \tag{24}$$

where $\mathbf{x} = [\mathbf{x}_1^T\ \mathbf{x}_2^T]^T$ and $\mathbf{y} = [\mathbf{y}_1^T\ \mathbf{y}_2^T]^T$, and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n, \mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m$. Then the corresponding alternative optimization is given by:

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{y}\|_2 \mathbf{h}^T\mathbf{x} + \|\mathbf{x}\|_2 \mathbf{f}^T\mathbf{y} + \psi(\mathbf{x}, \mathbf{y}), \tag{25}$$

where $\mathbf{h} \in \mathbb{R}^{2n}, \mathbf{f} \in \mathbb{R}^{2m}$ have i.i.d. standard Gaussian elements.

We note that this form is identical to that of the classical CGMT. A proof of this fact is given in the supplement section E.2.

5 APPLICATIONS

In this section, we discuss three applications of our comparison theorems. The first is regularized regression with vector-valued labels. As shown in the introduction, this problem is not accessible to the previously known comparison theorems. The second application is regression over complex variables. The third application is the case of regularized regression with convolutions. The third example shows a more complex form of weight sharing than vector-valued regression. Additional details regarding numerical simulations may be found in supplement section G.

5.1 Regularized Regression with Vector Valued Labels

We consider a dataset $\{(\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^m \times \mathbb{R}^K\}_{i=1}^n$, where $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We collect $\mathbf{x}_i$ and $\mathbf{y}_i$ into matrices $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $\mathbf{Y} \in \mathbb{R}^{n\times K}$ and consider a problem of the form:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{m\times K}} \frac{1}{2nK}\left\|\frac{1}{\sqrt{m}}\mathbf{X}\mathbf{\Theta}-\mathbf{Y}\right\|_F^2 + \frac{1}{mK}R(\mathbf{\Theta}). \quad (26)$$

Here $\|\cdot\|_F$ is the Frobenius norm, and $R(\mathbf{\Theta})$ is a strongly convex regularization function. Introducing the variable $\mathbf{Z}$ as the Legendre transform of the Frobenius norm, we obtain:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{m\times K}} \max_{\mathbf{Z}\in\mathbb{R}^{n\times K}} \frac{1}{nK\sqrt{m}}\mathrm{Tr}[\mathbf{Z}^T\mathbf{X}\mathbf{\Theta}] - \frac{1}{nk}\mathrm{Tr}[\mathbf{Z}^T\mathbf{Y}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{\Theta}). \quad (27)$$

Note that this is one of the special cases considered in corollary 1.

We assume a simple model for the labels, where $\mathbf{y}_i = (\mathbf{\Theta}^*)^T\mathbf{x}_i + \boldsymbol{\nu}_i$ where $\mathbf{\Theta}^* \in \mathbb{R}^{m\times K}$ is an underlying "true" model, and $\boldsymbol{\nu}_i \in \mathbb{R}^K$ are i.i.d. noise samples with zero mean and covariance $\mathbf{\Sigma} \in \mathbb{R}^{K\times K}$. By making use of our comparison theorem and simplifying the result, we can find the following alternative optimization problem:

$$\min_{\mathbf{Q}\in\mathcal{S}_+^K} \max_{\mathbf{P},\mathbf{V}\in\mathcal{S}_+^K} \frac{1}{K}\mathrm{Tr}[\mathbf{P}(\mathbf{Q}^T\mathbf{Q}+\mathbf{\Sigma})^{1/2}] - \frac{1}{2K}\mathrm{Tr}[\mathbf{P}^T\mathbf{P}+\mathbf{V}\mathbf{Q}^T\mathbf{Q}] - \frac{m}{2nK}\mathrm{Tr}[\mathbf{P}\mathbf{V}^{-1}\mathbf{P}] + \frac{1}{mK}\mathbb{E}_{\mathbf{F}}\mathcal{M}_{\mathbf{V}\ R(\cdot)}\left(\mathbf{\Theta}^* - \sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{P}\mathbf{V}^{-1}\right). \quad (28)$$

Here $\mathbf{F} \in \mathbb{R}^{m\times K}$ is an i.i.d. Gaussian matrix, and $\mathcal{M}$ is the matrix Moreau envelope over $R(\cdot)$.

The problem can be further simplified for a specific choice of the regularization function. We choose the case of quadratic regularization, where we choose $R(\mathbf{\Theta}) = \frac{1}{2}\mathrm{Tr}[\mathbf{\Theta}\mathbf{\Lambda}\mathbf{\Theta}^T]$, where $\mathbf{\Lambda} \in \mathbb{R}^{K\times K}$ is a PSD matrix of regularization parameters whose elements correspond to the pairwise regularization between classes. In this case we can simplify the alternative optimization further, to following form:

$$\min_{\mathbf{S}\in\mathcal{S}_+^K} \max_{\mathbf{T}\in\mathcal{S}_+^K} \frac{1}{2K}\mathrm{Tr}\left[\mathbf{\Lambda}\left(\frac{1}{m}\mathbf{\Theta}^{*T}\mathbf{\Theta}^* + \mathbf{S}^T\mathbf{S}\right)\right] - \frac{1}{2K}\mathrm{Tr}[\mathbf{T}^T\mathbf{T}] + \frac{1}{K}\mathrm{Tr}\left[\mathbf{T}\left(\mathbf{\Sigma}+\mathbf{S}^T\mathbf{S}\right)^{1/2}\right] - \frac{1}{K}\mathrm{Tr}\left[\mathbf{S}\left(\frac{1}{m}\mathbf{\Lambda}\mathbf{\Theta}^{*T}\mathbf{\Theta}^*\mathbf{\Lambda} + \frac{m}{n}\mathbf{T}^T\mathbf{T}\right)^{1/2}\right], \quad (29)$$

where the matrix square roots all denote the unique PSD square root for a PSD matrix. We also note that the resulting alternative optimization only contains variables of size $K \times K$. No element of the objective function grows with either n, m which allows for significantly more efficient computation of the alternative problem.

Furthermore, by theorem 3, we can also compute the expected generalization error. In figure 1 below we compare the theoretical prediction and numerical simulation of the generalization error of this problem for two choices of the number of classes $K = 3, 5$ and the regularization matrix $\mathbf{\Lambda}$ given by $\mathbf{\Lambda} = \lambda\mathbf{I}_K + 0.1\lambda(\mathbf{1}_{K\times K} - \mathbf{I}_K)$ for a base value of $\lambda \geq 0$, as a function of the ratio $\frac{m}{n}$. We observe that the theoretical values (lines) predicted by the CGMT-alternative optimization, and the numerical simulations (points) of the primary objective, match exactly. Furthermore, we observe that our theory is predicts a double descent phenomenon (Belkin et al., 2020) in this setup, which is suppressed as the regularization strength increases, as is common among regularized regression problems.

5.2 Regularized Complex Variable Regression

Next, we consider a complex-valued vector $\boldsymbol{\theta} \in \mathbb{C}^m$ of variables and a dataset $\{(\mathbf{z}_i, y_i) \in \mathbb{C}^m \times \mathbb{C}\}_{i=1}^n$, where $\mathbf{z}_i = \mathbf{a}_i + \mathrm{i}\mathbf{b}_i$ where both $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^m \sim \mathcal{N}(0, 1)$, and i denotes the imaginary unit. We collect $\mathbf{z}_i$ and y_i into a matrix $\mathbf{Z} \in \mathbb{C}^{n\times m}$ and vector $\mathbf{y} \in \mathbb{C}^n$, and consider a problem of the form:

$$\min_{\boldsymbol{\theta}\in\mathbb{C}^m} \frac{1}{4n}\left\|\frac{1}{\sqrt{2m}}\mathbf{Z}\boldsymbol{\theta}-\mathbf{y}\right\|_2^2 + \frac{1}{2m}R(\boldsymbol{\theta}), \quad (30)$$

where $\|\mathbf{a}\|_2 = \sqrt{\mathbf{a}^H\mathbf{a}}$, where ^{H} denotes the hermitian conjugate, and $R : \mathbb{C}^m \to \mathbb{R}$. We denote $\mathbf{Z} = \mathbf{G} +$

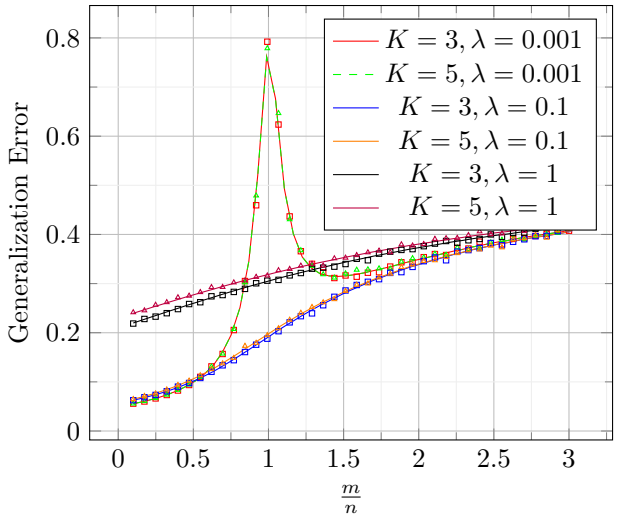

Figure 1: Generalization Error for Multi-class Regression for number of classes K and regularization $\mathbf{\Lambda} = \lambda \mathbf{I}_K + 0.1\lambda(\mathbf{1}_{K\times K} - \mathbf{I}_K)$ parameterized by λ as a function of the ratio of the number of parameters m to the number of data points n. Theory is given by solid lines and numerical simulations by triangles and squares.

$\mathrm{i}\mathbf{H}$, where $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{n\times m}$ are i.i.d. standard Gaussian matrices. Similarly, we denote $\boldsymbol{\theta} = \boldsymbol{\theta}_1 + \mathrm{i}\boldsymbol{\theta}_2$ and $\mathbf{y} = \mathbf{y}_1 + \mathrm{i}\mathbf{y}_2$. Then, we can express the problem in (30)as:

$$\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^m} \frac{1}{4n}\left\|\frac{1}{\sqrt{2m}}(\mathbf{G}\boldsymbol{\theta}_1 - \mathbf{H}\boldsymbol{\theta}_2) - \mathbf{y}_1\right\|^2 + \frac{1}{4n}\left\|\frac{1}{\sqrt{2m}}(\mathbf{G}\boldsymbol{\theta}_2 + \mathbf{H}\boldsymbol{\theta}_1) - \mathbf{y}_2\right\|^2 + \frac{1}{2m}R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2). \quad (31)$$

To proceed, we introduce $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^n$ as the Legendre transform of the two norms:

$$\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^m} \max_{\mathbf{z}_1,\mathbf{z}_2\in\mathbb{R}^n} \frac{1}{4nm}\begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{G} & -\mathbf{H}\\ \mathbf{H} & \mathbf{G}\end{bmatrix}\begin{bmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{bmatrix} + \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}^T\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} - \frac{1}{4n}\|\mathbf{z}_1\|^2 - \frac{1}{4n}\|\mathbf{z}_2\|^2 + \frac{1}{2m}R(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2). \quad (32)$$

This form is an optimization of the special case discussed in corollary 2.

Now, we assume a simple model for the labels $\mathbf{y}_i = \mathbf{z}_i^T\boldsymbol{\theta}^* + \nu_i$, where $\boldsymbol{\theta}^* \in \mathbb{C}^m$ is the "true" model, and $\nu_i \in \mathbb{C} = \nu_{i,1} + \mathrm{i}\nu_{i,2}$ is a zero mean complex noise, with noise power $\sigma^2_{\nu,1}, \sigma^2_{\nu,2}$ for $\nu_{i,1}, \nu_{i,2}$, respectively. The simplified alternative optimization can be expressed as:

$$\max_{\beta\geq 0}\min_{q\geq 0} \frac{\beta q}{2}\left(1 - \frac{m}{n}\right) + \frac{\beta(\sigma^2_{\boldsymbol{\nu},1} + \sigma^2_{\boldsymbol{\nu},2})}{4} - \frac{\beta^2}{2} + \frac{1}{4m}\mathbb{E}\mathcal{M}_{\frac{q}{\beta}\,R}\left(\boldsymbol{\theta}^* - \frac{q\sqrt{m}}{\sqrt{n}}\mathbf{f}\right). \quad (33)$$

where $\mathbf{f} \in \mathbb{R}^{2m}$ has i.i.d. standard Gaussian entries. In figure 2, we consider the case of R being the square loss $\frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$ and show the resulting theory for a variety of choices of λ as a function of the ratio $\frac{m}{n}$. Once again, we observe that the results match, and that double descent is observed in this setup around the interpolation threshold of $\frac{m}{n} = 1$.

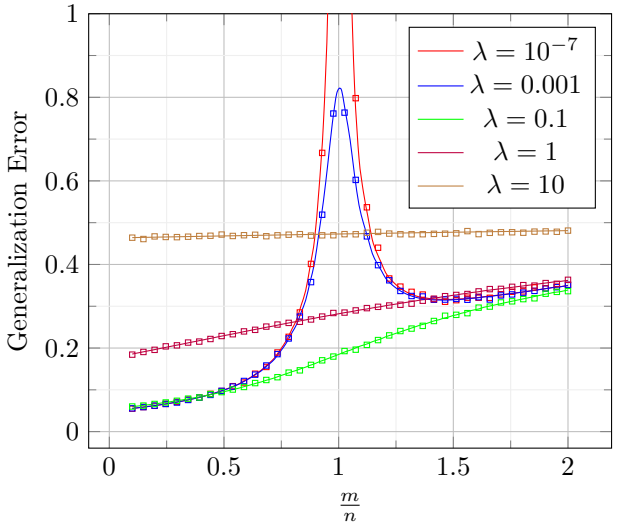

Figure 2: Generalization Error for Complex valued regression with square regularization for a number of choices of regularization strength λ as a function of the ratio between the number of parameters m and the number of data points n. Theory is given by lines and numerical simulations by squares.

5.3 Regularized Convolutional Regression

In this part, we study a single convolutional filter $\mathbf{\Theta} \in \mathbb{R}^{k_1\times k_2}$ and a data set $\{(\mathbf{X}_i, \mathbf{Y}_i) \in \mathbb{R}^{d_1\times d_2} \times \mathbb{R}^{D_1\times D_2}\}_{i=1}^n$ where $D_1 = d_1 - k_1 + 1$ and $D_2 = d_2 - k_2 + 1$. We require $d_1 \geq k_1, d_2 \geq k_2$, such that $D_1, D_2 \geq 1$. We consider the problem:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{k_1\times k_2}} \frac{1}{2nD_1D_2}\sum_{i=1}^n\left\|\frac{1}{\sqrt{k_1k_2}}\mathbf{X}_i * \mathbf{\Theta} - \mathbf{Y}_i\right\|_F^2 + \frac{1}{k_1k_2}R(\mathbf{\Theta}), \quad (34)$$

where $\|\cdot\|_F$ is the Frobenius norm, and $*$ is the (Machine Learning) convolution operator here between two matrices, and R is some strongly convex regularization function. We assume that each $\mathbf{X}_i$ has i.i.d. standard Gaussian entries. We first re-express the convolution as a matrix product. We define $\bar{\mathbf{X}} \in \mathbb{R}^{nD_1D_2\times k_1k_2}$

defined as

$$\left(\bar{\mathbf{X}}\right)_{\alpha D_1 D_2+\beta D_2+\gamma,\eta k_2+\epsilon} = \sum_{a=1}^{n}\sum_{b=1}^{d_1}\sum_{c=1}^{d_2}(\mathbf{X}_a)_{b,c}\delta_{\alpha,a}\delta_{\beta+\eta,b}\delta_{\gamma+\epsilon,c}, \quad (35)$$

Where $\alpha \in [n], \beta \in [D_1], \gamma \in [D_2], \eta \in [k_1], \epsilon \in [k_2]$. We further define $\boldsymbol{\theta} \in \mathbb{R}^{k_1 k_2} = \mathrm{vec}(\boldsymbol{\Theta})$ and $\mathbf{y} \in \mathbb{R}^{nD_1D_2}$ element-wise by $y_{\alpha D_1 D_2+\beta D_2+\gamma} = (\mathbf{Y}_\alpha)_{\beta,\gamma}$. Making use of the Legendre transform of the square loss, our original problem may be expressed as:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{k_1k_2}}\max_{\mathbf{z}\in\mathbb{R}^{nD_1D_2}} \frac{1}{nD_1D_2\sqrt{k_1k_2}}\mathbf{z}^T\bar{\mathbf{X}}\boldsymbol{\theta} - \frac{1}{nD_1D_2}\mathbf{z}^T\mathbf{y} - \frac{1}{2nD_1D_2}\|\mathbf{z}\|^2 + R(\mathrm{vec}^{-1}(\boldsymbol{\theta})). \quad (36)$$

Where we have introduced $\mathbf{z}$ by means of the Legendre transform of the ℓ_2^2-norm. In this case, the matrix $\bar{\mathbf{X}}$ is a GMS, and the bilinear form can be expressed as

$$\mathbf{z}^T\bar{\mathbf{X}}\boldsymbol{\theta} = \mathbf{z}^T\left(\sum_{\omega=1}^{k_1}\sum_{\nu=1}^{k_2}\mathbf{A}_{\omega\nu}^T\tilde{\mathbf{X}}\mathbf{b}_{\omega\nu}\right)\boldsymbol{\theta}, \quad (37)$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{nd_1d_2\times 1}$ has i.i.d. standard Gaussian entries, and $\mathbf{A}_{\omega\nu} \in \mathbb{R}^{nd_1d_2\times nD_1D_2}$ and $\mathbf{b}_{\omega\nu} \in \mathbb{R}^{1\times k_1k_2}$ are defined by:

$$\begin{aligned}(\mathbf{A}_{\omega\nu})_{\alpha D_1D_2+\beta D_2+\gamma,ad_1d_2+bd_2+c} &= \delta_{\alpha,a}\delta_{\omega+\beta,b}\delta_{\gamma+\nu,c},\\ (\mathbf{b}_{\omega\nu})_{\eta k_2+\epsilon} &= \delta_{\omega,\eta}\delta_{\nu,\epsilon}, \end{aligned} \quad (38)$$

where $a, \alpha \in [n], b \in [d_1], c \in [d_2], \beta \in [D_1], \gamma \in [D_2], \eta \in [k_1], \epsilon \in [k_2]$. Once again, we consider a simple model for the labels, we assume that $\mathbf{Y}_i = \mathbf{X}_i * \boldsymbol{\Theta}^* + \mathbf{N}_i$, where $\boldsymbol{\Theta}^* \in \mathbb{R}^{k_1\times k_2}$ is the "true" model and $\mathbf{N}_i \in \mathbb{R}^{D_1\times D_2}$ is a noise matrix with i.i.d. elements with zero mean and variance σ^2.

We can find the simplified alternative optimization corresponding to this problem, but due to its complexity we leave it in the supplement, see section F.3.1: Instead, as in the case of multiclass regression, the problem may be simplified further with a particular choice for the regularization function. Here we only consider the case of ridge regression, where we choose $R(\boldsymbol{\Theta}) = \frac{\lambda}{2}\|\boldsymbol{\Theta}\|_F^2$. In this case, the alternative optimization can be expressed as:

$$\begin{aligned}\min_{\mathbf{S}\in\mathcal{S}_+^{k_1k_2}}\max_{\mathbf{T}\in\mathcal{S}_+^{D_1D_2}} &\frac{\lambda}{2k_1k_2}\|\boldsymbol{\Theta}^*\|_F^2 - \frac{1}{2D_1D_2}\mathrm{Tr}\left[\mathbf{T}^T\mathbf{T}\right] + \frac{\lambda}{2k_1k_2}\mathrm{Tr}\left[\mathbf{S}^T\mathbf{S}\right]\\ &- \frac{1}{k_1k_2}\mathrm{Tr}\left[\mathbf{S}\left(\lambda^2\mathrm{vec}(\boldsymbol{\Theta}^*)\mathrm{vec}(\boldsymbol{\Theta}^*)^T + \frac{k_1k_2}{nD_1D_2}\mathbf{U}\right)^{1/2}\right]\\ &+ \frac{1}{D_1D_2}\mathrm{Tr}\left[\mathbf{T}\left(\sigma^2\mathbf{I}_{D_1D_2} + \mathbf{V}\right)^{1/2}\right],\end{aligned}$$

where $\mathbf{U} \in \mathbb{R}^{k_1k_2\times k_1k_2}$ and $\mathbf{V} \in \mathbb{R}^{D_1D_2\times D_1D_2}$ are defined by:

$$U_{\omega k_2+\nu,\omega' k_2+\nu} = \frac{1}{D_1D_2}\mathrm{Tr}\left[\mathbf{T}^T\mathbf{T}\tilde{\mathbf{A}}_{\omega\nu}^T\tilde{\mathbf{A}}_{\omega'\nu'}\right]$$

$$\mathbf{V} = \frac{1}{k_1k_2}\sum_{\omega,\omega'=1}^{k_1}\sum_{\nu,\nu'=1}^{k_2}(\mathbf{S}^T\mathbf{S})_{\omega k_2+\nu,\omega' k_2+\nu'}\tilde{\mathbf{A}}_{\omega\nu}^T\tilde{\mathbf{A}}_{\omega'\nu'}. \quad (39)$$

In figure 3 we consider the convolutional regression problem with square loss as a function of the number of data points n, for fixed choices $d_1 = d_2 = 20$ and $k_1 = k_2 = 9$. We again note that our theory accurately predicts the expected behavior of this setup. We can note that as n grows large the generalization error drops asymptotically to a minimal value, which for $\lambda = 0$ regularization is exactly the noise floor.

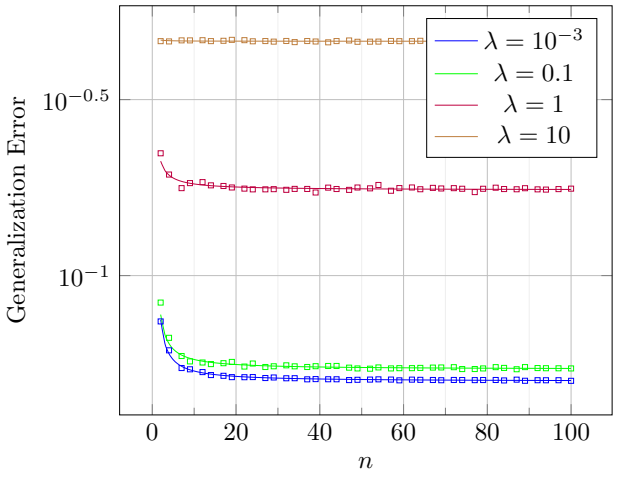

Figure 3: Generalization Error for Convolutional regression with square regularization for a number of choices of regularization strength λ as a function of the number of data points n. Theory is given by lines and numerical simulations by squares.

6 CONCLUSION

In this work, we addressed the key difficulty with the CGMT theorem, namely its requirement for i.i.d. Gaussian entries. We provided a generalized comparison theorem for min-max and max-max, convex optimization problems over Gaussian processes with GMS structure, and established equality of the expected optimal value and many other quantities of the optimal point, even in finite dimensions. This allowed us to obtain and verify exact learning curves for multiple problems with weight and feature sharing. Together with universality arguments, this result paves the path to the analysis of more complex models in future studies.

References

A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

D. Akhtiamov, R. Ghane, and B. Hassibi. Regularized linear regression for binary classification. *arXiv preprint arXiv:2311.02270*, 2023.

D. Akhtiamov, D. Bosch, R. Ghane, K. N. Varma, and B. Hassibi. A novel gaussian min-max theorem and its applications. *arXiv preprint arXiv:2402.07356*, 2024.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. 2013.

P. K. Andersen and R. D. Gill. Cox's regression model for counting processes: a large sample study. *The annals of statistics*, pages 1100–1120, 1982.

L. Aolaritei, S. Shafieezadeh-Abadeh, and F. Dörfler. The performance of wasserstein distributionally robust m-estimators in high dimensions. *arXiv:2206.13269*, 2023.

H. H. Bauschke, P. L. Combettes, et al. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

M. Belkin, D. Hsu, and J. Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, Jan 2020. ISSN 2577-0187. doi: 10.1137/20m1336072. URL `http://dx.doi.org/10.1137/20M1336072`.

D. Bosch, A. Panahi, A. Özcelikkale, and D. Dubhash. Double descent in random feature models: Precise asymptotic analysis for general convex regularization, 2022. URL `https://arxiv.org/abs/2204.02678`.

D. Bosch, A. Panahi, and B. Hassibi. Precise asymptotic analysis of deep random feature models. *arXiv preprint arXiv:2302.06210*, 2023.

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

S. Diamond and S. Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83): 1–5, 2016.

Y. Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50 (4):265–289, 1985.

Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

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

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

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

B. Loureiro, C. Gerbelot, H. Cui, S. Goldt, F. Krzakala, M. Mézard, and L. Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model, 2021. URL `https://arxiv.org/abs/2102.08127`.

F. Mignacco, F. Krzakala, Y. Lu, P. Urbani, and L. Zdeborova. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International Conference on Machine Learning*, pages 6874–6883. PMLR, 2020.

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

S. Oymak, C. Thrampoulidis, and B. Hassibi. The squared-error of generalized lasso: A precise analysis. In *2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1002–1009. IEEE, 2013.

A. Panahi and B. Hassibi. A universal analysis of large-scale regularized least squares solutions. In *NIPS*, pages 3384–3393, 2017.

M. Rudelson and R. Vershynin. Sparse reconstruction by convex relaxation: Fourier and gaussian measurements. In *2006 40th Annual Conference on Information Sciences and Systems*, pages 207–212. IEEE, 2006.

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

D. Slepian. The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2): 463–501, 1962.

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

M. Stojnic. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013b.

M. Stojnic. Fully bilinear generic and lifted random processes comparisons, 2016a. URL https://arxiv.org/abs/1612.08516.

M. Stojnic. Generic and lifted probabilistic comparisons – max replaces minmax, 2016b. URL https://arxiv.org/abs/1612.08506.

M. Stojnic. Bilinearly indexed random processes – *stationarization* of fully lifted interpolation, 2023a. URL https://arxiv.org/abs/2311.18097.

M. Stojnic. Fully lifted interpolating comparisons of bilinearly indexed random processes, 2023b. URL https://arxiv.org/abs/2311.18092.

M. Stojnic. Fully lifted random duality theory, 2023c. URL https://arxiv.org/abs/2312.00070.

C. Thrampoulidis, S. Oymak, and B. Hassibi. Simple error bounds for regularized noisy linear inverse problems. In *2014 IEEE International Symposium on Information Theory*, pages 3007–3011. IEEE, 2014.

C. Thrampoulidis, S. Oymak, and B. Hassibi. The Gaussian min-max theorem in the Presence of Convexity. *arXiv e-prints*, art. arXiv:1408.4837, Aug. 2014.

C. Thrampoulidis, A. Panahi, and B. Hassibi. Asymptotically exact error analysis for the generalized equation-lasso. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2021–2025. IEEE, 2015.

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

C. Thrampoulidis, S. Oymak, and M. Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view, 2020. URL https://arxiv.org/abs/2011.07729.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

L. Zhou, Z. Dai, F. Koehler, and N. Srebro. Uniform convergence with square-root lipschitz loss. *Advances in Neural Information Processing Systems*, 36, 2024.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Not Applicable]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Comparison Lemmas and Theorem's

In this section we state the currently existing comparison theorems that we extend, for the sake of completeness.

A.1 Slepian's Lemma and Slepian's Comparison Theorem

Theorem 4 (Slepian's Lemma (Slepian, 1962)). *Let* $\{ X_i \}_{i=1}^n$ *and* $\{ Y_i \}_{i=1}^n$ *be two sequences of real valued centered Gaussian random variables, which satisfy the following conditions:*

- $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2]$ *for all* $i \in [n]$
- $\mathbb{E}[X_i X_j] \leq E[Y_i Y_j]$ *for all* $i, j \neq i \in [n]$

Then for any $c_1, c_2, \ldots, c_n \in \mathbb{R}$*:*

$$\Pr\left[\bigcup_{i=1}^n X_i > c_i\right] \geq \Pr\left[\bigcup_{i=1}^n Y_i > c_i\right]. \tag{40}$$

There exist a standard pair of Gaussian processes that satisfy Slepian's Lemma which allows for the following max-max Theorem:

Theorem 5 (Slepian Max-Max Theorem). *Let* $\mathbf{G} \in \mathbb{R}^{n \times m}, \gamma \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ *have i.i.d. Standard Gaussian elements. Let* $\mathcal{X} \subset \mathbb{R}^n$ *and* $\mathcal{Y} \subset \mathbb{R}^m$ *be two compact sets and let* $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ *be a continuous function. Consider the following two Gaussian processes*

$$\begin{aligned} \mathcal{P}(\mathbf{G}, \boldsymbol{\gamma}) &= \max_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \mathbf{x}^T \mathbf{G} \mathbf{y} + \|\mathbf{x}\| \|\mathbf{y}\| \gamma + \psi(\mathbf{x}, \mathbf{y}), \\ \mathcal{A}(\mathbf{g}, \mathbf{h}) &= \max_{\mathbf{x} \in \mathcal{X}} \max_{\mathbf{y} \in \mathcal{Y}} \|\mathbf{x}\| \mathbf{g}^T \mathbf{y} + \|\mathbf{y}\| \mathbf{h}^T \mathbf{x} + \psi(\mathbf{x}, \mathbf{y}). \end{aligned} \tag{41}$$

Then, for any $c \in \mathbb{R}$

$$\Pr\left[\mathcal{P}(\mathbf{G}, \gamma) > c\right] \leq \Pr\left[\mathcal{A}(\mathbf{g}, \mathbf{h}) > c\right]. \tag{42}$$

A.2 Gordon's Lemma, Gaussian Min-Max Theorem, and the Convex Gaussian min-Max Theorem

Gordon (1985) later proved an extension of Slepian's comparisons lemma:

Theorem 6 (Gordon's Comparison Lemma (Gordon, 1985)). *Let* $\{ X_{i,j} \}$ *and* $\{ Y_{i,j} \}$ *for* $i \in [n]$ *and* $j \in [m]$ *be two sequences of real valued centered Gaussian random variables, which satisfy the following conditions:*

- $\mathbb{E}[X_{i,j}^2] = \mathbb{E}[Y_{i,j}^2]$ *for all* $i \in [n], j \in [m]$
- $\mathbb{E}[X_{i,j} X_{i,k}] \leq E[Y_{i,j} Y_{i,k}]$ *for all* $i \in [n], j, k \in [m]$
- $\mathbb{E}[X_{i,j} X_{l,k}] \geq E[Y_{i,j} Y_{l,k}]$ *for all* $i, l \neq i \in [n], j, k \in [m]$

Then, for any $c_{i,j} \in \mathbb{R}$ *for* $i \in [n], j \in [m]$*:*

$$\Pr\left[\bigcap_{i=1}^n \bigcup_{j=1}^m X_{i,j} \geq c_{i,j}\right] \geq \Pr\left[\bigcap_{i=1}^n \bigcup_{j=1}^m Y_{i,j} \geq c_{i,j}\right]. \tag{43}$$

Similarly to Slepian's Theorem the same set of processes hold for Gordon's Theorem, which is summarized in Gaussian Min-Max Theorem (GMT):

Theorem 7 (Gaussian Min-Max Theorem (GMT) (Gordon, 1985)). *Let $\mathbf{G} \in \mathbb{R}^{n\times m}, \gamma \in \mathbb{R}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ have i.i.d. Standard Gaussian elements. Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be two compact sets and let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Consider the following two Gaussian processes*

$$\begin{aligned}\mathcal{P}(\mathbf{G}, \boldsymbol{\gamma}) &= \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathbf{x}^T\mathbf{G}\mathbf{y} + \|\mathbf{x}\|\,\|\mathbf{y}\|\,\gamma + \psi(\mathbf{x},\mathbf{y}),\\ \mathcal{A}(\mathbf{g},\mathbf{h}) &= \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{x}\|\,\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|\,\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y}).\end{aligned} \tag{44}$$

Then, for any $c \in \mathbb{R}$

$$\Pr\left[\mathcal{P}(\mathbf{G},\gamma) \le c\right] \le \Pr\left[\mathcal{A}(\mathbf{g},\mathbf{h}) \le c\right]. \tag{45}$$

Later work by Thrampoulidis et al. (2014) extended this theorem to more general processes at the cost of additional convexity requirements on the sets $\mathcal{X}, \mathcal{Y}$ and function ψ.

Theorem 8 (Convex Gaussian Min-Max Theorem (CGMT) (Thrampoulidis et al., 2014)). *Let $\mathbf{G} \in \mathbb{R}^{n\times m}, \mathbf{g} \in \mathbb{R}^m, \mathbf{h} \in \mathbb{R}^n$ have i.i.d. Standard Gaussian elements and be independent of each other. Let $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ be two compact sets and let $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous function. Consider the following two Gaussian processes*

$$\begin{aligned}\mathcal{P}(\mathbf{G}, \boldsymbol{\gamma}) &= \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathbf{x}^T\mathbf{G}\mathbf{y} + \psi(\mathbf{x},\mathbf{y}),\\ \mathcal{A}(\mathbf{g},\mathbf{h}) &= \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \|\mathbf{x}\|\,\mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|\,\mathbf{h}^T\mathbf{x} + \psi(\mathbf{x},\mathbf{y}).\end{aligned} \tag{46}$$

Then, for any $c \in \mathbb{R}$

$$\Pr\left[\mathcal{P}(\mathbf{G}) \le c\right] \le 2\Pr\left[\mathcal{A}(\mathbf{g},\mathbf{h}) \le c\right]. \tag{47}$$

Furthermore, assume that $\mathcal{X}, \mathcal{Y}$ are convex sets, and that ψ is convex-concave on $\mathcal{X} \times \mathcal{Y}$, then for any $c_2 \in \mathbb{R}$:

$$\Pr\left[\mathcal{P}(\mathbf{G}) \ge c\right] \le 2\Pr\left[\mathcal{A}(\mathbf{g},\mathbf{h}) \ge c\right]. \tag{48}$$

B Proof of Theorem 1

B.1 Problem Setup

We consider two sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$ and consider a constant $K \in \mathbb{N}$. We consider a random matrix $\mathbf{G} \in \mathbb{R}^{n\times m}$ which is a Gaussian Matrix Sum, given by:

$$\mathbf{G} = \sum_{k=1}^{K} \mathbf{A}_k^T \tilde{\mathbf{G}} \mathbf{B}_k, \tag{49}$$

where $\mathbf{A}_k \in \mathbb{R}^{\tilde{n}\times n}, \mathbf{B}_k \in \mathbb{R}^{\tilde{m}\times m}$ are sets of deterministic matrices and $\tilde{\mathbf{G}} \in \mathbb{R}^{\tilde{n}\times\tilde{m}}$ is an i.i.d. standard Gaussian matrix. We further define the following two matrices $\mathbf{P}(\mathbf{x}), \mathbf{Q}(\mathbf{y}) \in \mathbb{R}^{K\times K}$ given by:

$$(\mathbf{P}(\mathbf{x}))_{k,k'} = \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x} \qquad (\mathbf{Q}(\mathbf{y}))_{k,k'} = \mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}. \tag{50}$$

We can note that $\mathbf{P}, \mathbf{Q}$ are Positive Semi-Definite (PSD) matrices, and therefore have canonical PSD square roots, which we denote as $\mathbf{P}^{1/2}(\mathbf{x})$ and $\mathbf{Q}^{1/2}(\mathbf{y})$ respectively.

We define the following two functions:

$$\begin{aligned}\mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi) = &\quad \mathbf{x}^T\left(\textstyle\sum_{k=1}^K \mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\right)\mathbf{y} + \mathrm{Tr}\left[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})\right] + \psi(\mathbf{x},\mathbf{y})\\ \mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi) = &\quad \textstyle\sum_{k=1}^K \mathbf{f}_k^T\mathbf{B}_k\mathbf{y} + \mathbf{h}_k^T\mathbf{A}_k\mathbf{x} + \psi(\mathbf{x},\mathbf{y}).\end{aligned} \tag{51}$$

Here $\psi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ and $\mathbf{F} \in \mathbb{R}^{\tilde{m}\times K}, \mathbf{H} \in \mathbb{R}^{\tilde{n}\times K}, \gamma \in \mathbb{R}^{K\times K}$ have i.i.d. standard Gaussian elements are independent of each other and independent of $\tilde{\mathbf{G}}$. The columns of $\mathbf{F}, \mathbf{H}$ are denoted by $\mathbf{f}_k$ and $\mathbf{h}_k$ respectively, and are defined as:

$$\mathbf{F} = \tilde{\mathbf{F}}\mathbf{P}^{1/2}(\mathbf{x}) \qquad \mathbf{H} = \tilde{\mathbf{H}}\mathbf{Q}^{1/2}(\mathbf{y}), \tag{52}$$

where $\tilde{\mathbf{F}} \in \mathbb{R}^{\tilde{m}\times K}, \tilde{\mathbf{H}} \in \mathbb{R}^{\tilde{n}\times K}$ have i.i.d. standard Gaussian elements. Finally, we recall the objects of interest, that being:

$$\mathcal{P}(\tilde{\mathbf{G}}, \boldsymbol{\gamma}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi) \qquad \mathcal{A}(\mathbf{F},\mathbf{H}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi). \tag{53}$$

B.2 Thermodynamic Representation

We consider a parameter $\beta > 0$, corresponding to the "inverse temperature", and a parameter $s \in \{ -1, 1 \}$. We consider the function

$$f(\tilde{\mathbf{G}}, \boldsymbol{\gamma}, \mathcal{X}, \mathcal{Y}, \beta, s, \psi) = \frac{1}{\beta |s|} \log \left(\int_{\mathcal{X}} d\mathbf{x} \left(\int_{\mathcal{Y}} d\mathbf{y} \ e^{\beta \mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \psi)} \right)^s \right). \tag{54}$$

Then, we see that:

$$\lim_{\beta \to \infty} \mathbb{E}_{\tilde{\mathbf{G}}, \boldsymbol{\gamma}} f(\tilde{\mathbf{G}}, \boldsymbol{\gamma}, \mathcal{X}, \mathcal{Y}, \beta, s, \psi) = \mathbb{E}_{\tilde{\mathbf{G}}, \boldsymbol{\gamma}} \max_{\mathbf{x} \in \mathcal{X}} \ s \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \psi). \tag{55}$$

Specializing to the case of $s = 1$ or $s = -1$, this results in a max-max or min-max problem respectively. We note that for this evaluation we have interchanged the limit over β and the expected value over $\mathbb{E}_{\tilde{\mathbf{G}}, \boldsymbol{\gamma}}$, this is justified by the dominated convergence theorem. We can first note that:

$$\lim_{\beta \to \infty} \frac{1}{\beta |s|} \log \left(\int_{\mathcal{X}} d\mathbf{x} \left(\int_{\mathcal{Y}} d\mathbf{y} \ e^{\beta \mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \psi)} \right)^s \right) = \max_{\mathbf{x} \in \mathcal{X}} s \max_{\mathbf{y} \in \mathcal{Y}} \mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \psi), \tag{56}$$

By the assumptions of strict convexity and/or concavity on ψ there exists a unique solution to the max-max or min-max optimization problem. Furthermore, by continuity of the objective this optimal point is not in a null set with respect to to the double integral. As such in the β limit the Boltzmann distribution will converge to the optimal value of $\mathcal{H}_{\mathcal{P}}$. Finally by the assumptions that the sets $\mathcal{X}$ and $\mathcal{Y}$ are compact there exists a maximum value of the logarithm independent of the choice of β, which justifies the usage of the dominated convergence theorem, to interchange the β limit and the expected value $\mathbb{E}_{\mathbf{G}, \boldsymbol{\gamma}}$.

To study the function $f(\tilde{\mathbf{G}}, \gamma, \mathcal{X}, \mathcal{Y}, \beta, s, \psi)$ we study the following interpolating function:

$$\xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, t) = \mathbb{E}_{\tilde{\mathbf{G}}, \boldsymbol{\gamma}, \mathbf{F}, \mathbf{H}} \frac{1}{\beta |s|} \log \left(\int_{\mathcal{X}} d\mathbf{x} \left(\int_{\mathcal{Y}} d\mathbf{y} \ e^{\beta \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)} \right)^s \right). \tag{57}$$

Here $t \in [0, 1]$ and $\mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)$ is given by:

$$\mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi) = \psi(\mathbf{x}, \mathbf{y}) + \sqrt{t} \left[\mathbf{x}^T \left(\sum_{k=1}^K \mathbf{A}_k^T \tilde{\mathbf{G}} \mathbf{B}_k \right) \mathbf{y} + \mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x}) \boldsymbol{\gamma} \mathbf{Q}^{1/2}(\mathbf{y})] \right] + \sqrt{1 - t} \left[\sum_{k=1}^K \mathbf{f}_k^T \mathbf{B}_k \mathbf{y} + \mathbf{h}_k^T \mathbf{A}_k \mathbf{x} \right]. \tag{58}$$

We can see therefore that $\xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, 1) = \mathbb{E}_{\mathbf{G}, \boldsymbol{\gamma}} f(\tilde{\mathbf{G}}, \boldsymbol{\gamma}, \mathcal{X}, \mathcal{Y}, \beta, s, \psi)$. For the sake of convenience we will define the set $\mathcal{U} = \{ \tilde{\mathbf{G}}, \boldsymbol{\gamma}, \mathbf{F}, \mathbf{H} \}$ and express ξ as:

$$\xi(\mathcal{X}, \mathcal{Y}, \beta, s, \psi, t) = \mathbb{E}_{\mathcal{U}} \frac{1}{\beta |s|} \log Z, \tag{59}$$

where:

$$Z \overset{def}{=} \int_{\mathcal{X}} d\mathbf{x} \ (C(\mathbf{x}))^s \qquad C(\mathbf{x}) \overset{def}{=} \int_{\mathcal{Y}} d\mathbf{y} \ A(\mathbf{x}, \mathbf{y}) \qquad A(\mathbf{x}, \mathbf{y}) \overset{def}{=} e^{\beta \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)}. \tag{60}$$

We will call Z the partition function.

To determine the properties of the interpolating function ξ we will examine its derivative with respect to t. We compute this in the following section. However we first prove some useful lemma and properties of this distribution. We first note that:

$$\Pi_1 = \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x}, \mathbf{y})}{Z} = \frac{\left(\int_{\mathcal{Y}} d\mathbf{y} e^{\beta \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)} \right)^{s-1} e^{\beta \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)}}{\int_{\mathcal{X}} d\mathbf{x} \left(\int_{\mathcal{Y}} d\mathbf{y} e^{\beta \mathcal{H}_t(\mathbf{x}, \mathbf{y}, \psi)} \right)^s}. \tag{61}$$

is a probability distribution, in the sense that $\int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x}, \mathbf{y})}{Z} d\mathbf{y} d\mathbf{x} = 1$. As such we define the following expected value with respect to this distribution:

$$\langle \cdot \rangle_{\Pi_1, \beta} = \int_{\mathbf{X}} d\mathbf{x} \int_{\mathcal{Y}} d\mathbf{y} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x}, \mathbf{y})}{Z} (\cdot). \tag{62}$$

Throughout the proof we will find two other similar probability distributions, which we denote by:

$$\Pi_2 = \frac{(C(\mathbf{x}))^{s-2} A(\mathbf{x},\mathbf{y}) A(\mathbf{x},\mathbf{y}')}{Z}, \qquad \Pi_3 = \frac{(C(\mathbf{x}))^{s-1}(C(\mathbf{x}'))^{s-1} A(\mathbf{x},\mathbf{y}) A(\mathbf{x}',\mathbf{y}')}{Z^2}. \tag{63}$$

we will denote the expected values with respect to these distributions by:

$$\begin{aligned} \langle\cdot\rangle_{\Pi_2,\beta} &= \frac{1}{Z}\int_{\mathcal{X}} d\mathbf{x}(C(\mathbf{x}))^{s-2}\int_{\mathcal{Y}} d\mathbf{y} A(\mathbf{x},\mathbf{y})\int_{\mathcal{Y}} d\mathbf{y}' A(\mathbf{x},\mathbf{y}')(\cdot), \\ \langle\cdot\rangle_{\Pi_3,\beta} &= \frac{1}{Z^2}\int_{\mathcal{X}} d\mathbf{x}(C(\mathbf{x}))^{s-1}\int_{\mathcal{X}} d\mathbf{x}'(C(\mathbf{x}'))^{s-1}\int_{\mathcal{Y}} d\mathbf{y} A(\mathbf{x},\mathbf{y})\int_{\mathcal{Y}} d\mathbf{y}' A(\mathbf{x}',\mathbf{y}')(\cdot). \end{aligned} \tag{64}$$

We now prove the following lemma:

Lemma 1. *Consider a continuous function $f : \mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$. Consider $\mathcal{H}_t$, and let $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ denote the unique optimal point of this objective (the point that corresponds to the max-max or min-max value; uniqueness guaranteed by the strict convexity/concavity). Then,*

$$\begin{aligned} \lim_{\beta\to\infty}\mathbb{E}_{\mathcal{U}}\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y})\rangle_{\Pi_1,\beta} &= \mathbb{E}_{\mathcal{U}} f(\hat{\mathbf{x}},\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{y}}) \\ \lim_{\beta\to\infty}\mathbb{E}_{\mathcal{U}}\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y}')\rangle_{\Pi_2,\beta} &= \mathbb{E}_{\mathcal{U}} f(\hat{\mathbf{x}},\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{y}}) \\ \lim_{\beta\to\infty}\mathbb{E}_{\mathcal{U}}\langle f(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\rangle_{\Pi_3,\beta} &= \mathbb{E}_{\mathcal{U}} f(\hat{\mathbf{x}},\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{y}}) \end{aligned} \tag{65}$$

Proof. We can first note that by the properties of the Boltzmann distribution and the fact that the optimal point is guarantee to be unique, we can see that the measures

$$\lim_{\beta\to\infty}\frac{A(\mathbf{x},\mathbf{y})}{C(\mathbf{x})} = \delta(\mathbf{y}-\hat{\mathbf{y}}(\mathbf{x})) \qquad \lim_{\beta\to\infty}\frac{(C(x))^s}{Z} = \delta(\mathbf{x}-\hat{\mathbf{x}}) \tag{66}$$

and that in the limit of $\beta\to\infty$ the product measure will converge to the product of the delta functions. As such, we note for $i = 1,2,3$ that:

$$\lim_{\beta\to\infty}\langle f(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\rangle_{\Pi_i,\beta} = f(\hat{\mathbf{x}},\hat{\mathbf{x}},\hat{\mathbf{y}},\hat{\mathbf{y}}), \tag{67}$$

by the properties of the Boltzmann distribution. This convergence is point-wise. As discussed above, the solution to the objective $\mathcal{H}_t$ is unique, and by the continuity of the objective $\mathcal{H}_t$ the unique solution is not in a null set of the integral. Finally, because of the fact that the integrals are bounded to compact sets and the fact that f is continuous implies the existence of a total upper bound M on the expected values independent of the choice of β, which by the dominated convergence theorem allows for the interchange of limits and expectation. □

As a consequence of this lemma we note the following corollary, which we will make use of frequently in the subsequent:

Corollary 3. Consider a continuous function $f : \mathcal{X}\times\mathcal{X}\times\mathcal{Y}\times\mathcal{Y}\to\mathbb{R}$, and let $s\in\{-1,1\}$ be a parameter, then:

$$\lim_{\beta\to\infty}\mathbb{E}_{\mathcal{U}}\left[\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y})\rangle_{\Pi_1,\beta} + (s-1)\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y}')\rangle_{\Pi_2,\beta} - s\langle f(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\rangle_{\Pi_3,\beta}\right] = 0, \tag{68}$$

and

$$\lim_{\beta\to\infty}\mathbb{E}_{\mathcal{U}}\beta\left[\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y})\rangle_{\Pi_1,\beta} + (s-1)\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y}')\rangle_{\Pi_2,\beta} - s\langle f(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\rangle_{\Pi_3,\beta}\right] = 0. \tag{69}$$

Proof. The first claim is directly justified by lemma 1. For the second claim we can note that:

$$\lim_{\beta\to\infty}\beta\left[\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y})\rangle_{\Pi_1,\beta} + (s-1)\langle f(\mathbf{x},\mathbf{x},\mathbf{y},\mathbf{y}')\rangle_{\Pi_2,\beta} - s\langle f(\mathbf{x},\mathbf{x}',\mathbf{y},\mathbf{y}')\rangle_{\Pi_3,\beta}\right] = 0. \tag{70}$$

This is due to the fact that the concentration of each $\langle f\rangle_{\Pi_i,\beta}$ on their optimal value is exponential. As such the sum concentrates to 0 at an exponential rate, overpowering the linear term. By the same argument as in lemma 1, the fact that f is continuous and $\mathbf{x},\mathbf{y}$ are restricted to compact sets allows for the use of the dominated convergence theorem, which gives the result. □

B.3 The t-derivative

We now compute the derivative of the function ξ with respect to t. We see that:

$$\begin{aligned}\frac{d\xi}{dt} &= \mathbb{E}_{\mathcal{U}} \frac{1}{\beta|s|} \frac{d}{dt} \log Z \\ &= \mathbb{E}_{\mathcal{U}} \frac{1}{\beta|s|Z} \frac{d}{dt} Z \\ &= \mathbb{E}_{\mathcal{U}} \frac{\beta s}{\beta|s|Z} \int_{\mathcal{X}} d\mathbf{x} (C(\mathbf{x}))^{s-1} \int_{\mathcal{Y}} d\mathbf{y}\ A(\mathbf{x},\mathbf{y}) \frac{d}{dt} \mathcal{H}_t. \end{aligned} \tag{71}$$

We can see that

$$\begin{aligned}\frac{d\mathcal{H}_t}{dt} = \frac{\sum_{k=1}^{K} \mathbf{x}^T \mathbf{A}_k^T \tilde{\mathbf{G}} \mathbf{B}_k \mathbf{y} + \sum_{k,k',k''} \gamma_{k',k''} (\mathbf{P}^{1/2}(\mathbf{x}))_{k,k'} (\mathbf{Q}^{1/2}(\mathbf{y}))_{k'',k}}{2\sqrt{t}} \\ - \frac{\sum_{k,k'=1}^{K} \sum_{i=1}^{\tilde{m}} \sum_{j=1}^{m} (\mathbf{P}^{1/2}(\mathbf{x}))_{k,k'} F_{i,k'} (\mathbf{B}_k)_{i,j} y_j + \sum_{k,k'=1}^{K} \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{n} (\mathbf{Q}^{1/2}(\mathbf{y}))_{k,k'} H_{i,k'} (\mathbf{A}_k)_{i,j} x_j}{2\sqrt{1-t}}. \end{aligned} \tag{72}$$

We can the express the derivative with respect to t as

$$\frac{d\xi}{dt} = \frac{\text{sign}(s)}{2} \int_{\mathcal{X}} d\mathbf{x} \int_{\mathcal{Y}} d\mathbf{y}\ \sum_{k=1}^{K} \left(T_{\tilde{\mathbf{G}},k} + \sum_{k'=1}^{K} T_{\boldsymbol{\gamma},k,k'} - T_{\mathbf{F},k} - T_{\mathbf{H},k} \right), \tag{73}$$

where

$$\begin{aligned} T_{\tilde{\mathbf{G}},k} &= \frac{1}{\sqrt{t}} \sum_{a=1}^{\tilde{n}} \sum_{b=1}^{\tilde{m}} \mathbb{E}_{\mathcal{U}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{A}_k \mathbf{x})_a (\mathbf{B}_k \mathbf{y})_b \tilde{G}_{a,b}}{Z}, \\ T_{\boldsymbol{\gamma},k,k'} &= \frac{1}{\sqrt{t}} \mathbb{E}_{\mathcal{U}} \sum_{k''=1}^{K} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{P}^{1/2}(\mathbf{x}))_{k'',k} (\mathbf{Q}^{1/2}(\mathbf{y}))_{k'',k'} \gamma_{k,k'}}{Z}, \\ T_{\mathbf{F},k} &= \frac{1}{\sqrt{1-t}} \sum_{k'=1}^{K} \sum_{b=1}^{\tilde{m}} \mathbb{E}_{\mathcal{U}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{P}^{1/2}(\mathbf{x}))_{k,k'} (\mathbf{B}_{k'} \mathbf{y})_b F_{b,k}}{Z}, \\ T_{\mathbf{H},k} &= \frac{1}{\sqrt{1-t}} \sum_{k'=1}^{K} \sum_{a=1}^{\tilde{n}} \mathbb{E}_{\mathcal{U}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{Q}^{1/2}(\mathbf{y}))_{k,k'} (\mathbf{A}_{k'} \mathbf{x})_a H_{a,k}}{Z}. \end{aligned} \tag{74}$$

We now examine each of these terms in turn.

B.3.1 The $T_{\tilde{\mathbf{G}},k}$ group

We can first make use of Gaussian integration by parts to fnd the following:

$$\begin{aligned} T_{\tilde{\mathbf{G}},k} &= \frac{1}{\sqrt{t}} \sum_{a=1}^{\tilde{n}} \sum_{b=1}^{\tilde{m}} \mathbb{E}_{\mathcal{U}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{A}_k \mathbf{x})_a (\mathbf{B}_k \mathbf{y})_b \tilde{G}_{a,b}}{Z} \\ &= \frac{1}{\sqrt{t}} \sum_{a=1}^{\tilde{n}} \sum_{b=1}^{\tilde{m}} \mathbb{E}_{\mathcal{U}} \left[\sum_{a'=1}^{\tilde{m}} \sum_{b'=1}^{\tilde{n}} \mathbb{E}_{\mathcal{U}} (\tilde{G}_{a,b} \tilde{G}_{a',b'}) \frac{d}{d\tilde{G}_{a',b'}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{A}_k \mathbf{x})_a (\mathbf{B}_k \mathbf{y})_b}{Z} \right] \\ &= \frac{1}{\sqrt{t}} \sum_{a=1}^{\tilde{n}} \sum_{b=1}^{\tilde{m}} \mathbb{E}_{\mathcal{U}} \left[\frac{d}{d\tilde{G}_{a,b}} \frac{(C(\mathbf{x}))^{s-1} A(\mathbf{x},\mathbf{y}) (\mathbf{A}_k \mathbf{x})_a (\mathbf{B}_k \mathbf{y})_b}{Z} \right]. \end{aligned} \tag{75}$$

Where we have made use of the fact that $\tilde{\mathbf{G}}$ has i.i.d. elements, as such $\mathbb{E}_{\mathcal{U}}\tilde{G}_{ab}\tilde{G}_{a'b'} = \delta_{aa'}\delta_{bb'}$. We can then see that:

$$T_{\tilde{\mathbf{G}},k} = \sum_{a=1}^{\tilde{n}}\sum_{b=1}^{\tilde{m}}\left[\mathbb{E}_{\mathcal{U}}\frac{\beta(C(\mathbf{x}))^{s-1}A(\mathbf{x},\mathbf{y})\sum_{k'=1}^{K}(\mathbf{A}_k\mathbf{x})_a(\mathbf{A}_{k'}\mathbf{x})_a(\mathbf{B}_k\mathbf{y})_b(\mathbf{B}_{k'}\mathbf{y})_b}{Z}\right.$$
$$+\mathbb{E}_{\mathcal{U}}\frac{\beta(s-1)(C(\mathbf{x}))^{s-2}A(\mathbf{x},\mathbf{y})\int_{\mathcal{Y}}d\mathbf{y}'\ A(\mathbf{x},\mathbf{y}')\sum_{k'=1}^{K}(\mathbf{A}_k\mathbf{x})_a(\mathbf{A}_{k'}\mathbf{x})_a(\mathbf{B}_k\mathbf{y})_b(\mathbf{B}_{k'}\mathbf{y}')_b}{Z}$$
$$\left.-\mathbb{E}_{\mathcal{U}}\frac{\beta s(C(\mathbf{x}))^{s-1}A(\mathbf{x},\mathbf{y})\int_{\mathcal{X}}d\mathbf{x}'\ (C(\mathbf{x}'))^{s-1}\int_{\mathcal{Y}}d\mathbf{y}'\ A(\mathbf{x}',\mathbf{y}')\sum_{k'=1}^{K}(\mathbf{A}_k\mathbf{x})_a(\mathbf{A}_{k'}\mathbf{x}')_a(\mathbf{B}_k\mathbf{y})_b(\mathbf{B}_{k'}\mathbf{y}')_b}{Z^2}\right] \tag{76}$$

From this we can see that:

$$\int_{\mathcal{X}}d\mathbf{x}\int_{\mathcal{Y}}d\mathbf{y}\sum_{k=1}^{K}T_{\tilde{G},k} = \beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}\right.$$
$$\left.+(s-1)\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\right\rangle_{\Pi_2,\beta} - s\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}'\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\right\rangle_{\Pi_3,\beta}\right]. \tag{77}$$

B.3.2 The Other Groups

The same strategy of Gaussian integration by parts can be applied to re-express the other terms. We can find that for $T_{\boldsymbol{\gamma},k,k'}$ that:

$$\int_{\mathcal{X}}d\mathbf{x}\int_{\mathcal{Y}}d\mathbf{y}\sum_{k,k'=1}^{K}T_{\boldsymbol{\gamma},k,k'} = \beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k'}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}(\mathbf{y}))_{k,k'}\right\rangle_{\Pi_1,\beta}\right.$$
$$\left.+(s-1)\left\langle\sum_{k,k'}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\right\rangle_{\Pi_2,\beta} - s\left\langle\sum_{k,k'}(\mathbf{P}^{1/2}(\mathbf{x})\mathbf{P}^{1/2}(\mathbf{x}'))_{k,k'}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\right\rangle_{\Pi_3,\beta}\right]. \tag{78}$$

For $T_{\mathbf{F},k}$ we find that:

$$\int_{\mathcal{X}}d\mathbf{x}\int_{\mathcal{Y}}d\mathbf{y}\sum_{k=1}^{K}T_{\mathbf{F},k} = \beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}\right.$$
$$\left.+(s-1)\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\right\rangle_{\Pi_2,\beta} - s\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}^{1/2}(\mathbf{x})\mathbf{P}^{1/2}(\mathbf{x}'))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\right\rangle_{\Pi_3,\beta}\right]. \tag{79}$$

and finally for $T_{\mathbf{H},k}$ we find that

$$\int_{\mathcal{X}}d\mathbf{x}\int_{\mathcal{Y}}d\mathbf{y}\sum_{k=1}^{K}T_{\mathbf{H},k} = \beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k'=1}^{K}(\mathbf{Q}(\mathbf{y}))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta}\right.$$
$$\left.+(s-1)\left\langle\sum_{k,k'=1}^{K}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_2,\beta} - s\left\langle\sum_{k,k'=1}^{K}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}'\right\rangle_{\Pi_3,\beta}\right]. \tag{80}$$

We can see that all 4 of these groups take the form specified in corollary 3, from which we can conclude the following theorem.

Theorem 9. *Consider the function $\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,t)$ and the setup described in section B.1. Then*

$$\lim_{\beta\to\infty}\frac{d\xi}{dt} = 0. \tag{81}$$

This gives rise to the obvious corollary:

Corollary 4. Assuming the setup of theorem 9, then we have that:

$$\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,t)=\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,0)+\int_0^t \frac{d\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,\tau)}{dt}d\tau, \tag{82}$$

and we obtain the following comparison principle:

$$\lim_{\beta\to\infty}\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,0)=\lim_{\beta\to\infty}\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,t)=\lim_{\beta\to\infty}\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,1). \tag{83}$$

B.4 Return to the Original Problem

Now, we recall that:

$$\lim_{\beta\to\infty}\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,1)=\mathbb{E}_{\tilde{\mathbf{G}},\gamma}\max_{\mathbf{x}\in\mathcal{X}} s\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi), \tag{84}$$

and similarly

$$\lim_{\beta\to\infty}\xi(\mathcal{X},\mathcal{Y},\beta,s,\psi,0)=\mathbb{E}_{\mathbf{g},\mathbf{h}}\max_{\mathbf{x}\in\mathcal{X}} s\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi). \tag{85}$$

As such by corollary 4 we see that:

$$\mathbb{E}_{\tilde{\mathbf{G}},\boldsymbol{\gamma}}\max_{\mathbf{x}\in\mathcal{X}} s\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi)=\mathbb{E}_{\mathbf{F},\mathbf{H}}\max_{\mathbf{x}\in\mathcal{X}} s\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi). \tag{86}$$

when specializing to the two different cases of $s=1$ and $s=-1$ we obtain the following results for $\max-\max$ problems:

$$\mathbb{E}_{\tilde{\mathbf{G}},\boldsymbol{\gamma}}\max_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi)=\mathbb{E}_{\mathbf{F},\mathbf{H}}\max_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi). \tag{87}$$

and the following result for $\min-\max$ problems:

$$\mathbb{E}_{\tilde{\mathbf{G}},\boldsymbol{\gamma}}\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{P}}(\mathbf{x},\mathbf{y},\psi)=\mathbb{E}_{\mathbf{F},\mathbf{H}}\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\mathcal{H}_{\mathcal{A}}(\mathbf{x},\mathbf{y},\psi). \tag{88}$$

This completes the first claim of the proof.

B.5 Relating the Variance

Now that we have demonstrated the relationship between the expected values of the two optimizations, we also with to compare their variance. To study this we consider the following object:

$$\mathbb{E}_{\mathcal{U}}\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\right\rangle_{\Pi_1,\beta} \tag{89}$$

which in the large β limit will evaluate to $\mathbb{E}_{\mathcal{U}}[\mathcal{H}_t^2(\hat{\mathbf{x}},\hat{\mathbf{y}},\psi)]$ evaluated at the optimal point of the max-max or min-max optimization $(\hat{\mathbf{x}},\hat{\mathbf{y}})$ by lemma 1. We will compute the t-derivative of this quantity:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{U}}\frac{d}{dt}\frac{1}{Z}\int_{\mathcal{X}}d\mathbf{x}(C(\mathbf{x}))^{s-1}\int_{\mathcal{Y}}d\mathbf{y}A(\mathbf{x},\mathbf{y})\mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\\
=\mathbb{E}_{\mathcal{U}}\beta\Bigg[&\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\frac{d\mathcal{H}_t(\mathbf{x},\mathbf{y},\psi)}{dt}\right\rangle_{\Pi_1,\beta}+(s-1)\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\frac{d\mathcal{H}_t(\mathbf{x},\mathbf{y}',\psi)}{dt}\right\rangle_{\Pi_2,\beta}\\
&-s\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\frac{d\mathcal{H}_t(\mathbf{x}',\mathbf{y}',\psi)}{dt}\right\rangle_{\Pi_3,\beta}\Bigg]+2\mathbb{E}_{\mathcal{U}}\left\langle \mathcal{H}_t(\mathbf{x},\mathbf{y})\frac{d\mathcal{H}_t(\mathbf{x},\mathbf{y},\psi)}{dt}\right\rangle_{\Pi_1,\beta}\\
&=2\mathbb{E}_{\mathcal{U}}\left\langle \mathcal{H}_t(\mathbf{x},\mathbf{y})\frac{d\mathcal{H}_t(\mathbf{x},\mathbf{y},\psi)}{dt}\right\rangle_{\Pi_1,\beta}+C_\beta.
\end{aligned} \tag{90}$$

We note that the first 3 terms are of the form of corollary 3, and will therefore go to 0 in the large β limit. We collect all these terms in a constant C_β, where $\lim_{\beta\to\infty} C_\beta = 0$. We can then note that:

$$\begin{aligned}
&\mathbb{E}_{\mathcal{U}}\left\langle \mathcal{H}_t(\mathbf{x},\mathbf{y})\frac{d\mathcal{H}_t(\mathbf{x},\mathbf{y},\psi)}{dt}\right\rangle_{\Pi_1,\beta} = \\
&\frac{1}{2}\mathbb{E}_{\mathcal{U}}\left\langle \left(\sum_{k=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}+\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right)^2-\left(\sum_{k=1}^{K}\mathbf{f}_k^T\mathbf{B}_k\mathbf{y}+\mathbf{h}_k^T\mathbf{A}_k\mathbf{x}\right)^2\right. \\
&\left.+\left(\frac{\sqrt{1-t}}{\sqrt{t}}-\frac{\sqrt{t}}{\sqrt{1-t}}\right)\left(\sum_{k=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}+\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right)\left(\sum_{k=1}^{K}\mathbf{f}_k^T\mathbf{B}_k\mathbf{y}+\mathbf{h}_k^T\mathbf{A}_k\mathbf{x}\right)\right\rangle_{\Pi_1,\beta}. \qquad (91)
\end{aligned}$$

We can now use the same strategy of using Gaussian integration by parts to examine each of these terms in turn:

$$\begin{aligned}
\mathbb{E}_{\mathcal{U}}\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}\mathbf{x}^T\mathbf{A}_{k'}\tilde{\mathbf{G}}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta} &= \mathbb{E}_{\mathcal{U}}\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta} \\
&+\sqrt{t}\beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k',k''=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}\mathbf{x}^T\mathbf{A}_{k'}\tilde{\mathbf{G}}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}\right. \\
&+(s-1)\left\langle\sum_{k,k',k''=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}'\mathbf{x}^T\mathbf{A}_{k'}\tilde{\mathbf{G}}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_2,\beta} \\
&\left.-s\left\langle\sum_{k,k',k''=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}'\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}'\mathbf{x}^T\mathbf{A}_{k'}\tilde{\mathbf{G}}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_3,\beta}\right] \\
&= \mathbb{E}_{\mathcal{U}}\left\langle\sum_{k,k'=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}+\sqrt{t}C_\beta, \qquad (92)
\end{aligned}$$

and,

$$\begin{aligned}
\mathbb{E}_{\mathcal{U}}\left\langle\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_1,\beta} &= \mathbb{E}_{\mathcal{U}}\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}(\mathbf{y}))_{k,k'}\right\rangle_{\Pi_1,\beta} \\
&+\sqrt{t}\beta\mathbb{E}_{\mathcal{U}}\left[\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}(\mathbf{y}))_{k,k'}\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_1,\beta}\right. \\
&+(s-1)\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_2,\beta} \\
&\left.-s\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}^{1/2}(\mathbf{x})\mathbf{P}^{1/2}(\mathbf{x}'))_{k,k'}(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_3,\beta}\right] \\
&= \mathbb{E}_{\mathcal{U}}\left\langle\sum_{k,k'=1}^{K}(\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}(\mathbf{y}))_{k,k'}\right\rangle_{\Pi_1,\beta}+\sqrt{t}C_\beta, \qquad (93)
\end{aligned}$$

and,

$$\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} \mathbf{f}_k^T\mathbf{B}_k\mathbf{y}\mathbf{f}_{k'}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta} = \mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} (\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}$$
$$+\sqrt{1-t}\beta\mathbb{E}_{\mathcal{U}}\left[\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{P}(\mathbf{x}))_{k,k''}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}\mathbf{f}_{k'}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta}\right.$$
$$+(s-1)\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{P}(\mathbf{x}))_{k,k''}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}'\mathbf{f}_{k'}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_2,\beta}$$
$$\left.-s\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{P}^{1/2}(\mathbf{x})\mathbf{P}^{1/2}(\mathbf{x}'))_{k,k''}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k''}\mathbf{y}\mathbf{f}_{k'}\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_3,\beta}\right]$$
$$= \mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} (\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta} + \sqrt{1-t}C_\beta, \tag{94}$$

and,

$$\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} \mathbf{h}_k^T\mathbf{A}_k\mathbf{x}\mathbf{h}_{k'}\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta} = \mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} (\mathbf{Q}(\mathbf{y}))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta}$$
$$+\sqrt{1-t}\beta\mathbb{E}_{\mathcal{U}}\left[\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{Q}(\mathbf{y}))_{k,k''}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}\mathbf{h}_{k'}\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta}\right.$$
$$+(s-1)\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k''}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}\mathbf{h}_{k'}\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_2,\beta}$$
$$\left.-s\left\langle \sum_{k,k',k''=1}^{K} (\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k''}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k''}\mathbf{x}'\mathbf{h}_{k'}\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_3,\beta}\right]$$
$$= \mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} (\mathbf{Q}(\mathbf{y}))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta} + \sqrt{1-t}C_\beta, \tag{95}$$

and,

$$\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_1,\beta} =$$
$$\sqrt{t}\beta\mathbb{E}_{\mathcal{U}}\left[\left\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_1,\beta}\right.$$
$$+(s-1)\left\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_2,\beta}$$
$$\left.-s\left\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}'\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}'\mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\right\rangle_{\Pi_3,\beta}\right] = \sqrt{t}C_\beta. \tag{96}$$

We also see that:

$$\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} \mathbf{f}_k^T\mathbf{B}_k\mathbf{y}\mathbf{h}_{k'}^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta} = \sqrt{1-t}C_\beta, \tag{97}$$

and for the cross terms

$$\begin{aligned}
\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}\mathbf{f}_{k'}^T\mathbf{B}_{k'}\mathbf{y}\right\rangle_{\Pi_1,\beta} &= \sqrt{1-t}C_\beta = \sqrt{t}\tilde{C}_\beta,\\
\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{x}\mathbf{h}_{k'}^T\mathbf{A}_{k'}\mathbf{x}\right\rangle_{\Pi_1,\beta} &= \sqrt{1-t}C_\beta = \sqrt{t}\tilde{C}_\beta,\\
\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k=1}^{K} \mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\mathbf{y}\mathbf{f}_k^T\mathbf{B}_k\mathbf{y}\right\rangle_{\Pi_1,\beta} &= \sqrt{1-t}C_\beta = \sqrt{t}\tilde{C}_\beta,\\
\mathbb{E}_{\mathcal{U}}\left\langle \sum_{k=1}^{K} \mathrm{Tr}[\mathbf{P}^{1/2}(\mathbf{x})\boldsymbol{\gamma}\mathbf{Q}^{1/2}(\mathbf{y})]\mathbf{h}_k^T\mathbf{A}_k\mathbf{x}\right\rangle_{\Pi_1,\beta} &= \sqrt{1-t}C_\beta = \sqrt{t}\tilde{C}_\beta,
\end{aligned} \tag{98}$$

either has a factor $\sqrt{1-t}$ or a factor $\sqrt{t}$ depending on around which Gaussian you complete the integration by parts. Collecting all the terms, and ensuring that all factors of $\sqrt{1-t}$ and $\sqrt{t}$ in the denominators are cancelled we can find that:

$$\begin{aligned}
\mathbb{E}_{\mathcal{U}}\frac{d}{dt}\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\right\rangle_{\Pi_1,\beta} = \mathbb{E}_{\mathcal{U}}\left\langle \sum_{k,k'=1}^{K}\left(\mathbf{x}^T\mathbf{A}_k\mathbf{A}_{k'}\mathbf{x} - (\mathbf{P}(\mathbf{x}))_{k,k'}\right)\left(\mathbf{y}^T\mathbf{B}_k\mathbf{B}_{k'}\mathbf{y} - (\mathbf{Q}(\mathbf{y}))_{k,k'}\right)\right\rangle_{\Pi_1,\beta} + C_\beta\\
= C_\beta,
\end{aligned} \tag{99}$$

where in the final equality we have made use of the definitions of $\mathbf{P}(\mathbf{x})$ and $\mathbf{Q}(\mathbf{y})$. As such, making use of the dominated convergence theorem we can find that:

$$\lim_{\beta\to\infty}\frac{d}{dt}\mathbb{E}_{\mathcal{U}}\left\langle \mathcal{H}_t^2(\mathbf{x},\mathbf{y},\psi)\right\rangle_{\Pi_1,\beta} = 0 \tag{100}$$

Then letting $(\hat{\mathbf{x}}_{\mathcal{P}},\hat{\mathbf{y}}_{\mathcal{P}})$ and $(\hat{\mathbf{x}}_{\mathcal{A}},\hat{\mathbf{y}}_{\mathcal{A}})$ be the optimal points of $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{A}}$ respectively, we can conclude that: From which we can conclude that:

$$\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\mathcal{H}_{\mathcal{P}}^2(\hat{\mathbf{x}}_{\mathcal{P}},\hat{\mathbf{y}}_{\mathcal{P}},\psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}}\mathcal{H}_{\mathcal{A}}^2(\hat{\mathbf{x}}_{\mathcal{A}},\hat{\mathbf{y}}_{\mathcal{A}},\psi). \tag{101}$$

Which combined with the results of part 1 of the theorem, allows us to find that:

$$\mathrm{Var}_{\mathbf{G},\boldsymbol{\gamma}}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}_{\mathcal{P}},\hat{\mathbf{y}}_{\mathcal{P}},\psi) = \mathrm{Var}_{\mathbf{F},\mathbf{H}}\mathcal{H}_{\mathcal{A}}(\hat{\mathbf{x}}_{\mathcal{A}},\hat{\mathbf{y}}_{\mathcal{A}},\psi). \tag{102}$$

C Proof of Theorem 2

This proof proceeds in much the same manner as the proof in section B. In this case we consider the primary optimization given by:

$$\mathcal{H}_{\mathcal{R}}(\mathbf{x},\mathbf{y},\psi) = \quad \mathbf{x}^T\left(\textstyle\sum_{k=1}^{K}\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\right)\mathbf{y} + \psi(\mathbf{x},\mathbf{y}). \tag{103}$$

We define a similar function $\tilde{\xi}$ defined by

$$\tilde{\xi}(\mathcal{X},\mathcal{Y},\beta,s,\psi,t) = \mathbb{E}_{\tilde{\mathbf{G}},\mathbf{F},\mathbf{H}}\frac{1}{\beta|s|}\log\left(\int_{\mathcal{X}}d\mathbf{x}\left(\int_{\mathcal{Y}}d\mathbf{y}\ e^{\beta\tilde{\mathcal{H}}_t(\mathbf{x},\mathbf{y},\psi)}\right)^s\right), \tag{104}$$

where

$$\tilde{\mathcal{H}}_t(\mathbf{x},\mathbf{y},\psi) = \psi(\mathbf{x},\mathbf{y}) + \sqrt{t}\left[\sum_{k=1}^{K}\mathbf{x}^T\mathbf{A}_k^T\tilde{\mathbf{G}}\mathbf{B}_k\mathbf{y}\right] + \sqrt{1-t}\left[\sum_{k=1}^{K}\mathbf{f}_k^T\mathbf{B}_k\mathbf{y} + \mathbf{h}_k^T\mathbf{A}_k\mathbf{x}\right]. \tag{105}$$

We can compute the t-derivative of $\tilde{\xi}$, using the same methods as described in section B.3, and we can find the following:

$$\frac{d\tilde{\xi}}{dt} = \frac{\beta \text{ sign}(s)}{2}\mathbb{E}_{\mathcal{U}}\Bigg[\Bigg\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y} - (\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y} \\ -(\mathbf{Q}(\mathbf{y}))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\Big\rangle_{\Pi_1,\beta} \\ +(s-1)\Bigg\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}' - (\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}' \\ -(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\Big\rangle_{\Pi_2,\beta} \\ -s\Bigg\langle \sum_{k,k'=1}^{K} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}'\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}' - (\mathbf{P}^{1/2}(\mathbf{x})\mathbf{P}^{1/2}(\mathbf{x}'))_{k,k'}\mathbf{y}^T\mathbf{B}_k^T\mathbf{B}_{k'}\mathbf{y}' \\ -(\mathbf{Q}^{1/2}(\mathbf{y})\mathbf{Q}^{1/2}(\mathbf{y}'))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}'\Big\rangle_{\Pi_3,\beta}\Bigg]. \tag{106}$$

We can note that all terms are the same, except we do not have the terms that relate to $\boldsymbol{\gamma}$. By corollary 3, we can similarly find that

$$\lim_{\beta\to\infty} \frac{d\tilde{\xi}}{dt} = 0. \tag{107}$$

The same arguments as in section B.4, result in the following two results. In the case of $s = 1$ and $\psi(\mathbf{x}, \mathbf{y})$ is strictly concave in both $\mathbf{x}$ and $\mathbf{y}$

$$\mathbb{E}_{\tilde{\mathbf{G}}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} H_{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \max_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} H_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \psi), \tag{108}$$

and similarly in the case where $s = -1$ and $\psi(\mathbf{x}, \mathbf{y})$ is strictly convex in $\mathbf{x}$ and strictly concave in $\mathbf{y}$.

$$\mathbb{E}_{\tilde{\mathbf{G}}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} H_{\mathcal{R}}(\mathbf{x}, \mathbf{y}, \psi) = \mathbb{E}_{\mathbf{F},\mathbf{H}} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} H_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \psi). \tag{109}$$

Which completes the first claim of the proof.

C.1 Relating the Variances

The proof for comparing the variances is completed identically to the one discussed in section B.5. All terms remain the same except we are now missing terms with respect to $\boldsymbol{\gamma}$. Collecting all of the terms we can find the following:

$$\frac{d}{dt}\mathbb{E}_{\mathcal{U}} \langle \tilde{\mathcal{H}}_t^2(\mathbf{x}, \mathbf{y}, \psi)\rangle_{\Pi_1,\beta} \\ = \mathbb{E}_{\mathcal{U}}\Bigg\langle \sum_{k,k'} \mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\mathbf{y}^T\mathbf{B}_k\mathbf{B}_{k'}\mathbf{y} - (\mathbf{P}(\mathbf{x}))_{k,k'}\mathbf{y}^T\mathbf{B}_k\mathbf{B}_{k'}\mathbf{y} - (\mathbf{Q}(\mathbf{y}))_{k,k'}\mathbf{x}^T\mathbf{A}_k^T\mathbf{A}_{k'}\mathbf{x}\Bigg\rangle_{\Pi_1,\beta} + C_\beta \\ = -\mathbb{E}_{\mathcal{U}}\Bigg\langle \sum_{k,k'} (\mathbf{P}(\mathbf{x}))_{k,k'}(\mathbf{Q}(\mathbf{y}))_{k,k'}\Bigg\rangle_{\Pi_1,\beta} + C_\beta \\ = -\mathbb{E}_{\mathcal{U}} \langle \text{Tr}[\mathbf{P}(\mathbf{x})\mathbf{Q}(\mathbf{y})]\rangle_{\Pi_1,\beta} + C_\beta. \tag{110}$$

Where in the third line we made use of the definition of $\mathbf{P}(\mathbf{x})$ and $\mathbf{Q}(\mathbf{y})$. We can now note that $\mathbf{P}(\mathbf{x})$ and $\mathbf{Q}(\mathbf{x})$ are PSD matrices, and the trace of the product of two PSD matrices is always a positive number. As we can therefore conclude that:

$$\lim_{\beta\to\infty} \frac{d}{dt}\mathbb{E}_{\mathcal{U}} \langle \tilde{\mathcal{H}}_t^2(\mathbf{x}, \mathbf{y}, \psi)\rangle_{\Pi_1,\beta} \leq 0. \tag{111}$$

Then letting $(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}})$ and $(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}})$ be the optimal points of $\mathcal{H}_{\mathcal{R}}$ and $\mathcal{H}_{\mathcal{A}}$ respectively, we can conclude that:

$$\mathbb{E}_{\mathbf{G}}\mathcal{H}^2_{\mathcal{R}}(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}}, \psi) \leq \mathbb{E}_{\mathbf{F},\mathbf{H}}\mathcal{H}^2_{\mathcal{A}}(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}}, \psi). \tag{112}$$

Which combined with the results of the first part of this theorem allows us to conclude that:

$$\mathrm{Var}_{\mathbf{G}}\mathcal{H}_{\mathcal{R}}(\hat{\mathbf{x}}_{\mathcal{R}}, \hat{\mathbf{y}}_{\mathcal{R}}, \psi) \leq \mathrm{Var}_{\mathbf{F},\mathbf{H}}\mathcal{H}_{\mathcal{A}}(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}}, \psi). \tag{113}$$

This concludes the proof.

D Proof of Theorem 3

To prove this statement, we consider a small value of $\epsilon > 0$, and instead consider the function $\psi_\epsilon(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x}, \mathbf{y}) + \epsilon\phi(\mathbf{x}, \mathbf{y})$. Provided that ϵ is sufficently small $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{A}}$ remain strictly concave in $\mathbf{y}$. We then denote by $(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}})$ as the optimal point of $\mathcal{H}_{\mathcal{P}}(\mathbf{x}, \mathbf{y}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon)$ and similarly $(\hat{\mathbf{x}}^\epsilon_{\mathcal{A}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{A}})$ as the optimal point of $\mathcal{H}_{\mathcal{A}}(\mathbf{x}, \mathbf{y}, \mathbf{F}, \mathbf{H}, \psi_\epsilon)$.

We can note that because $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{A}}$ are finite, and continuous, due to the continuity of ψ and ϕ, for all $\mathbf{x}$ and $\mathbf{y}$ in some small neighborhood around the optimal point the value of $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{A}}$ remains finite. Provided that ϵ is sufficiently small, and by the fact that ϕ is bounded, $\mathcal{H}_{\mathcal{P}}$ and $\mathcal{H}_{\mathcal{A}}$ evaluated at the points $(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}})$ and $(\hat{\mathbf{x}}^\epsilon_{\mathcal{A}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{A}})$ respectively will also be finite.

We can then see that:

$$\phi(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}}) = \frac{d}{d\epsilon}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon)\Big|_{\epsilon=0} = \lim_{\epsilon\to 0}\frac{\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon) - \mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi)}{\epsilon}. \tag{114}$$

We can then further note that:

$$\frac{d}{d\epsilon}\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon) = \mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\frac{d}{d\epsilon}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon). \tag{115}$$

This may be readily observed by the dominated convergence theorem theorem by the fact that $\phi(\mathbf{x}, \mathbf{y})$ can be upper bounded by $\|\phi\|_\infty$ for all $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, in other words we can see that

$$\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon) \leq \mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi + \epsilon\|\phi\|_\infty) < \infty. \tag{116}$$

As such we can then finally conclude that

$$\begin{aligned}\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\phi(\hat{\mathbf{x}}_{\mathcal{P}}, \hat{\mathbf{y}}_{\mathcal{P}}) = \mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\frac{d}{d\epsilon}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\mathbb{E}_{\mathbf{G},\boldsymbol{\gamma}}\mathcal{H}_{\mathcal{P}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{P}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{P}}, \mathbf{G}, \boldsymbol{\gamma}, \psi_\epsilon)\Big|_{\epsilon=0} \\ &= \frac{d}{d\epsilon}\mathbb{E}_{\mathbf{F},\mathbf{H}}\mathcal{H}_{\mathcal{A}}(\hat{\mathbf{x}}^\epsilon_{\mathcal{A}}, \hat{\mathbf{y}}^\epsilon_{\mathcal{A}}, \mathbf{F}, \mathbf{H}, \psi_\epsilon)\Big|_{\epsilon=0} = \mathbb{E}_{\mathbf{F},\mathbf{H}}\phi(\hat{\mathbf{x}}_{\mathcal{A}}, \hat{\mathbf{y}}_{\mathcal{A}}).\end{aligned} \tag{117}$$

The same argument may be repeated for the second claim concerning $\mathcal{H}_{\mathcal{R}}$.

E Proof of Corollaries 1 and 2

E.1 Proof of Corollary 1

We consider the trace form of interest in this case $\mathrm{Tr}[\mathbf{X}^T\mathbf{G}\mathbf{Y}]$, where $\mathbf{X} \in \mathbb{R}^{n\times k}$, $\mathbf{Y} \in \mathbb{R}^{m\times k}$ and $\mathbf{G} \in \mathbb{R}^{n\times m}$ with i.i.d. standard Gaussian entries. We note that we can express this form as a GMS, as follows:

$$\mathrm{Tr}[\mathbf{X}^T\mathbf{G}\mathbf{Y}] = \mathbf{x}^T\left(\sum_{l=1}^{k}\mathbf{A}_l^T\mathbf{G}\mathbf{B}_l\right)\mathbf{y}, \tag{118}$$

where $\mathbf{x} \in \mathbb{R}^{nk} = \mathrm{vec}(\mathbf{X})$ and $\mathbf{y} \in \mathbb{R}^{mk} = \mathrm{vec}(\mathbf{Y})$, and $\mathbf{A}_l \in \mathbb{R}^{nk\times n}, \mathbf{B}_l \in \mathbb{R}^{mk\times m}$ are given by:

$$(\mathbf{A}_l)_{an+b,c} = \delta_{a,l}\delta_{b,c} \qquad (\mathbf{B}_l)_{am+d,e} = \delta_{a,l}\delta_{d,e}, \tag{119}$$

where $a \in [K], b, c \in [n], d, e \in [m]$. We can see that in this case that the elements of $(\mathbf{P}(\mathbf{x}))_{l,l'}$ are given by:

$$(\mathbf{P}(\mathbf{x}))_{l,l'} = \mathbf{x}^T \mathbf{A}_l^T \mathbf{A}_{l'} \mathbf{x} = \sum_{a,a'=1}^{k} \sum_{b,b',c=1}^{n} \mathbf{x}_{an+b} \delta_{a,l} \delta_{b,c} \delta_{a',l} \delta_{b',c} \mathbf{x}_{a'n+b'} = \sum_{b=1}^{n} \mathbf{x}_{ln+b} \mathbf{x}_{l'n+b}. \tag{120}$$

Or equivalently we can see that $\mathbf{P}(\mathbf{x}) \in \mathbb{R}^{k \times k} = \mathbf{X}^T \mathbf{X}$. Similarly we can see that $\mathbf{Q}(\mathbf{y}) \in \mathbb{R}^{k \times k} = \mathbf{Y}^T \mathbf{Y}$. We can then note that by Theorem 2 that the first terms of the alternative optimization are given by:

$$\begin{aligned}
&\sum_{l=1}^{k} \mathbf{f}_l^T \mathbf{B}_l^T \mathbf{y} + \mathbf{h}_l^T \mathbf{A}_l \mathbf{x} \\
&= \sum_{a,l=1}^{k} \sum_{b,c=1}^{n} \sum_{d,e=1}^{m} \mathbf{F}_{l,d} (\mathbf{B}_l)_{d,an+e} \mathbf{y}_{an+e} + \mathbf{H}_{l,b} (\mathbf{A}_l)_{b,an+c} \mathbf{x}_{an+c} \\
&= \sum_{a,l,l'=1}^{k} \sum_{b,c=1}^{n} \sum_{d,e=1}^{m} (\mathbf{P}^{1/2})_{l,l'} \tilde{\mathbf{F}}_{l',d} (\mathbf{B}_l)_{d,an+e} \mathbf{y}_{an+e} + (\mathbf{Q}^{1/2})_{l,l'} \tilde{\mathbf{H}}_{l',b} (\mathbf{A}_l)_{b,an+c} \mathbf{x}_{an+c} \\
&= \sum_{l,l'=1}^{k} \sum_{b}^{n} \sum_{d}^{m} (\mathbf{P}^{1/2})_{l,l'} \tilde{\mathbf{F}}_{l',d} \mathbf{y}_{ln+d} + (\mathbf{Q}^{1/2})_{l,l'} \tilde{\mathbf{H}}_{l',b} \mathbf{x}_{ln+b} \\
&= \mathrm{Tr}[(\mathbf{X}^T \mathbf{X})^{1/2} \tilde{\mathbf{F}} \mathbf{Y}] + \mathrm{Tr}[(\mathbf{Y}^T \mathbf{Y})^{1/2} \tilde{\mathbf{H}} \mathbf{X}].
\end{aligned} \tag{121}$$

Where in the second line we made use of the definitions of $\mathbf{F}, \mathbf{H}$ with columns $\mathbf{f}_l, \mathbf{h}_l$ respectively. In the third line we made use of the definition of $\mathbf{F} = \tilde{\mathbf{F}} \mathbf{P}^{1/2}$ and $\mathbf{H} = \tilde{\mathbf{H}} \mathbf{Q}^{1/2}$, where $\tilde{\mathbf{F}} \in \mathbb{R}^{mk \times k}$ and $\tilde{\mathbf{H}} \in \mathbb{R}^{nk \times k}$ have i.i.d. standard Gaussian entries. In the fourth line we made use of the definitions of $\mathbf{A}_l$ and $\mathbf{B}_l$, and in the final line the definitions of $\mathbf{P}$ and $\mathbf{Q}$.

E.2 Proof of Corollary 2

We consider the complex bilinear form of interest for this special case:

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}^T \begin{bmatrix} \mathbf{G} & -\mathbf{H} \\ \mathbf{H} & \mathbf{G} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \tag{122}$$

where $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$, $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m$ and $\mathbf{G}, \mathbf{H} \in \mathbb{R}^{n \times m}$ are i.i.d. standard Gaussian and independent. We can first note that the Gaussian matrix can be expressed as a GMS:

$$\begin{bmatrix} \mathbf{G} & -\mathbf{H} \\ \mathbf{H} & \mathbf{G} \end{bmatrix} = \underbrace{\begin{bmatrix} \mathbf{I}_n & \mathbf{0}_n \\ \mathbf{0}_n & \mathbf{I}_n \end{bmatrix}}_{\mathbf{A}_1} \begin{bmatrix} \mathbf{G} \\ \mathbf{H} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{I}_m & \mathbf{0}_m \end{bmatrix}}_{\mathbf{B}_1} + \underbrace{\begin{bmatrix} \mathbf{0}_n & -\mathbf{I}_n \\ \mathbf{I}_n & \mathbf{0}_n \end{bmatrix}}_{\mathbf{A}_2} \begin{bmatrix} \mathbf{G} \\ \mathbf{H} \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{0}_m & \mathbf{I}_m \end{bmatrix}}_{\mathbf{B}_2}. \tag{123}$$

We can see that in this case that $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{2 \times 2}$ are given by:

$$\mathbf{P}(\mathbf{x}) = \begin{bmatrix} \|\mathbf{x}\|^2 & 0 \\ 0 & \|\mathbf{x}\|^2 \end{bmatrix} \qquad \mathbf{Q}(\mathbf{y}) = \begin{bmatrix} \|\mathbf{y}_1\|^2 & \mathbf{y}_1^T \mathbf{y}_2 \\ \mathbf{y}_1^T \mathbf{y}_2 & \|\mathbf{y}_2\|^2 \end{bmatrix}. \tag{124}$$

Where $\mathbf{x} = [\mathbf{x}_1^T \ \mathbf{x}_2^T]^T$, and similarly we will define $\mathbf{y} = [\mathbf{y}_1^T \ \mathbf{y}_2^T]^T$. We can then see that the alternative optimization take the form

$$\underbrace{\left(\mathbf{f}_1^T \mathbf{B}_1^T + \mathbf{f}_2^T \mathbf{B}_2^T\right)}_{\mathbf{f}} \mathbf{y} + \underbrace{\left(\mathbf{h}_1^T \mathbf{A}_1^T + \mathbf{h}_2^T \mathbf{A}_2^T\right)}_{\mathbf{h}} \mathbf{x}. \tag{125}$$

We can note that $\mathbf{f}$ is zero mean and has covariance:

$$\mathbb{E}[\mathbf{f}\mathbf{f}^T] = \sum_{i,i'=1}^{2} P_{i,i'} \mathbf{B}_i \mathbf{B}_{i'}^T = \|\mathbf{x}\|^2 \mathbf{I}_{2m}, \tag{126}$$

and similary $\mathbf{h}$ is zero mean and has covariance:

$$\mathbb{E}[\mathbf{h}\mathbf{h}^T] = \sum_{i,i'=1}^{2} Q_{i,i'}\mathbf{A}_i\mathbf{A}_{i'}^T = \begin{bmatrix} (\|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2)\mathbf{I}_n & (\mathbf{y}_1^T\mathbf{y}_2 - \mathbf{y}_1\mathbf{y}_2)\mathbf{I}_n \\ (\mathbf{y}_1^T\mathbf{y}_2 - \mathbf{y}_1^T\mathbf{y}_2)\mathbf{I}_n & (\|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2)\mathbf{I}_n \end{bmatrix} = \|\mathbf{y}\|^2 \mathbf{I}_{2n}. \tag{127}$$

As such we can finally conclude that the alternative optimization will take the form:

$$\|\mathbf{x}\|_2 \tilde{\mathbf{f}}^T\mathbf{y} + \|\mathbf{y}\|_2 \tilde{\mathbf{h}}^T\mathbf{x}, \tag{128}$$

where $\tilde{\mathbf{f}} \in \mathbb{R}^{2m}$ and $\tilde{\mathbf{h}} \in \mathbb{R}^{2n}$ have i.i.d. Gaussian entries.

F Applications

In this section we consider the applications discussed in section 5, where we apply our CGMT extention and simply the results. We take a similar approach to simplification of the alternatives as (Loureiro et al., 2021; Bosch et al., 2023).

F.1 Vector Valued Regression

We recall the optimization problem of interest in this case:

$$\min_{\mathbf{\Theta}\in\mathbb{R}^{m\times k}} \max_{\mathbf{Z}\in\mathbb{R}^{n\times k}} \frac{1}{nK\sqrt{m}}\mathrm{Tr}[\mathbf{Z}^T\mathbf{X}\mathbf{\Theta}] - \frac{1}{nK}\mathrm{Tr}[\mathbf{Z}^T\mathbf{Y}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{\Theta}). \tag{129}$$

We will assume for simplicity that R is strongly convex. We then further recall that we choose the following model for the labels $\mathbf{Y} = \mathbf{X}\mathbf{\Theta}^* + \mathbf{N}$, where $\mathbf{N} \in \mathbb{R}^{n\times K}$ is zero mean noise, with $\frac{1}{n}\mathbb{E}\mathbf{N}^T\mathbf{N} = \mathbf{\Sigma}$. We therefore obtain:

$$\min_{\mathbf{E}\in\mathbb{R}^{m\times k}} \max_{\mathbf{Z}\in\mathbb{R}^{n\times k}} \frac{1}{nK\sqrt{m}}\mathrm{Tr}[\mathbf{Z}^T\mathbf{X}\mathbf{E}] - \frac{1}{nk}\mathrm{Tr}[\mathbf{Z}^T\mathbf{N}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{E} + \mathbf{\Theta}^*). \tag{130}$$

where $\mathbf{E} = \mathbf{\Theta} - \mathbf{\Theta}^*$. We first show that this problem can be restricted to compact and convex sets:

Lemma 2. *Let $\hat{\mathbf{E}}, \hat{\mathbf{Z}}$ be the optimal values of* (130)*. Then there exists positive constants $C_{\mathbf{E}}$ and $C_{\mathbf{Z}}$ such that:*

$$\Pr\left(\left\|\hat{\mathbf{E}}\right\|_2 \le C_{\mathbf{E}}\sqrt{m}\right) \xrightarrow[m\to\infty]{P} 1, \qquad \Pr\left(\left\|\hat{\mathbf{Z}}\right\|_2 \le C_{\mathbf{Z}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1. \tag{131}$$

Proof. We can first note that by the strong convexity of the optimization problem with respect to $\mathbf{E}$ that the set of minimizers is a compact set, as such there there exists some constant $C_{\mathbf{E}}$ that is independent of m, such that:

$$\left\|\hat{\mathbf{E}}\right\|_2 \le C_{\mathbf{E}}\sqrt{m}. \tag{132}$$

We can next consider the optimality condition for $\mathbf{Z}$, taking the derivative we can find:

$$\hat{\mathbf{Z}} = \frac{1}{\sqrt{m}}\mathbf{X}\mathbf{E} - \mathbf{N}, \tag{133}$$

we can therefore see that

$$\|\mathbf{Z}\|_2 \le \frac{1}{\sqrt{m}}\|\mathbf{X}\|_2\|\mathbf{E}\|_2 + \|\mathbf{N}\|_2. \tag{134}$$

We can note that $\mathbf{X}$ and $\mathbf{N}$ are random matrices, whose operator norms are bounded $C\sqrt{n}$ in the limit of large n, for some constant C by classical results in random matrix theory (Vershynin, 2018). We further know that $\|\mathbf{E}\| \le C_{\mathbf{E}}\sqrt{m}$ with high probability, as such there exists some constant $C_{\mathbf{Z}}$ independent of n such that

$$\Pr(\|\mathbf{Z}\|_2 \le C_{\mathbf{Z}}\sqrt{n}) \xrightarrow[n\to\infty]{P} 1. \tag{135}$$

□

As such we can construct sets $\mathcal{S}_{\mathbf{E}} = \{ \mathbf{E} \in \mathbb{E}^{m\times K} \mid \|\mathbf{E}\|_2 \leq C_{\mathbf{E}}\sqrt{m} \}$ and $\mathcal{S}_{\mathbf{Z}} = \{ \mathbf{Z} \in \mathbb{E}^{n\times K} \mid \|\mathbf{Z}\|_2 \leq C_{\mathbf{Z}}\sqrt{n} \}$ and reduce our optimization to the following form:

$$\min_{\mathbf{E}\in\mathcal{S}_{\mathbf{E}}} \max_{\mathbf{Z}\in\mathcal{S}_{\mathbf{Z}}} \frac{1}{nK\sqrt{m}}\text{Tr}[\mathbf{Z}^T\mathbf{X}\mathbf{E}] - \frac{1}{nk}\text{Tr}[\mathbf{Z}^T\mathbf{N}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{E}+\mathbf{\Theta}^*). \tag{136}$$

We can now apply the special case of our novel CGMT extention to this setup to obtain the following optimization:

$$\min_{\mathbf{E}\in\mathcal{S}_{\mathbf{E}}} \max_{\mathbf{Z}\in\mathcal{S}_{\mathbf{Z}}} \frac{1}{nK\sqrt{m}}\text{Tr}[(\mathbf{Z}^T\mathbf{Z})^{1/2}\mathbf{F}\mathbf{E}] + \frac{1}{nK\sqrt{m}}\text{Tr}[(\mathbf{E}^T\mathbf{E})^{1/2}\mathbf{H}\mathbf{Z}] - \frac{1}{nk}\text{Tr}[\mathbf{Z}^T\mathbf{N}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{\Theta}). \tag{137}$$

We now let $\mathbf{A} = \frac{1}{\sqrt{n}}(\mathbf{Z}^T\mathbf{Z})^{1/2}$ and $\mathbf{B} = \frac{1}{\sqrt{m}}(\mathbf{E}^T\mathbf{E})^{1/2}$ and we reintroduce these constraints using Lagrange multipliers $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$:

$$\begin{aligned}\min_{\mathbf{E},\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{Z},\mathbf{A},\bar{\mathbf{B}}} \frac{1}{K\sqrt{mn}}\text{Tr}[\mathbf{A}\mathbf{F}\mathbf{E}] + \frac{1}{nK}\text{Tr}[\mathbf{B}\mathbf{H}\mathbf{Z}] - \frac{1}{nk}\text{Tr}[\mathbf{Z}^T\mathbf{N}] - \frac{1}{2nK}\|\mathbf{Z}\|_F^2 + \frac{1}{mK}R(\mathbf{E}+\mathbf{\Theta}^*)\\ +\frac{1}{2K}\text{Tr}[\bar{\mathbf{A}}(\mathbf{A}^T\mathbf{A} - \frac{1}{n}\mathbf{Z}^T\mathbf{Z})] + \frac{1}{2K}\text{Tr}[\bar{\mathbf{B}}(\frac{1}{m}\mathbf{E}^T\mathbf{E} - \mathbf{B}^T\mathbf{B})].\end{aligned} \tag{138}$$

We can then solve over $\mathbf{Z}$ to find

$$\mathbf{Z} = (\mathbf{I}+\bar{\mathbf{A}})^{-1}\left(\mathbf{H}^T\mathbf{B} - \mathbf{N}\right). \tag{139}$$

Substituting we obtain:

$$\begin{aligned}\min_{\mathbf{E},\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2K}\text{Tr}[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A} - \bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{K\sqrt{mn}}\text{Tr}[\mathbf{A}\mathbf{F}\mathbf{E}] + \frac{1}{2mK}\text{Tr}[\bar{\mathbf{B}}\mathbf{E}^T\mathbf{E}] + \frac{1}{mK}R(\mathbf{E}+\mathbf{\Theta}^*)\\ +\frac{1}{2nK}\text{Tr}[(\mathbf{H}^T\mathbf{B}-\mathbf{N})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}(\mathbf{H}^T\mathbf{B}-\mathbf{N})].\end{aligned} \tag{140}$$

We can then complete the square over $\mathbf{E}$ to find:

$$\begin{aligned}\min_{\mathbf{E},\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2K}\text{Tr}[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A} - \bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{mK}\text{Tr}[\left(\mathbf{E} - \sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right)^T\bar{\mathbf{B}}\left(\mathbf{E}^T - \sqrt{\frac{m}{n}}\bar{\mathbf{B}}^{-1}\mathbf{A}\mathbf{F}\right)]\\ +\frac{1}{mK}R(\mathbf{E}+\mathbf{\Theta}^*) - \frac{1}{2nK}\text{Tr}[\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}\mathbf{F}] + \frac{1}{2nK}\text{Tr}[(\mathbf{H}^T\mathbf{B}-\mathbf{N})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}(\mathbf{H}^T\mathbf{B}-\mathbf{N})].\end{aligned} \tag{141}$$

We can reintroduce $\mathbf{\Theta} = \mathbf{E} + \mathbf{\Theta}^*$ and recognize the Moreau envelope over R to find:

$$\begin{aligned}\min_{\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2K}\text{Tr}[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A} - \bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{mK}\mathcal{M}_{\bar{\mathbf{B}}\ R}\left(\mathbf{\Theta}^* - \sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right)\\ -\frac{1}{2nK}\text{Tr}[\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}\mathbf{F}] + \frac{1}{2nK}\text{Tr}[(\mathbf{H}^T\mathbf{B}-\mathbf{N})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}(\mathbf{H}^T\mathbf{B}-\mathbf{N})].\end{aligned} \tag{142}$$

Then in the following lemma we show that this result concentrates for large values of m, n.

Lemma 3. *Consider the term:*

$$\begin{aligned}F(\mathbf{A},\bar{\mathbf{A}},\mathbf{B},\bar{\mathbf{B}}) = \frac{1}{2K}Tr[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A} - \bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{mK}\mathcal{M}_{\bar{\mathbf{B}}\ R}\left(\mathbf{\Theta}^* - \sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right)\\ -\frac{1}{2nK}Tr[\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}\mathbf{F}] + \frac{1}{2nK}Tr[(\mathbf{H}^T\mathbf{B}-\mathbf{N})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}(\mathbf{H}^T\mathbf{B}-\mathbf{N})],\end{aligned} \tag{143}$$

and let $\bar{F}$ be given by

$$\begin{aligned}\bar{F}(\mathbf{A},\bar{\mathbf{A}},\mathbf{B},\bar{\mathbf{B}}) = \frac{1}{2K}Tr[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A} - \bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{mK}\mathbb{E}\mathcal{M}_{\bar{\mathbf{B}}\ R}\left(\mathbf{\Theta}^* - \sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right)\\ -\frac{m}{2nK}Tr[\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}] + \frac{1}{2K}Tr[(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}].\end{aligned} \tag{144}$$

Then

$$\min_{\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{A},\bar{\mathbf{B}}} F \xrightarrow[n,m\to\infty]{P} \min_{\mathbf{B},\bar{\mathbf{A}}} \max_{\mathbf{A},\bar{\mathbf{B}}} \bar{F}. \tag{145}$$

Proof. We can see clearly from standard results in probably theory (Vershynin, 2018) that

$$\frac{1}{n}\text{Tr}[(\mathbf{H}^T\mathbf{B}-\mathbf{N})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}(\mathbf{H}^T\mathbf{B}-\mathbf{N})], \tag{146}$$

will converge to

$$\text{Tr}[(\mathbf{B}^T\mathbf{B}+\mathbf{\Sigma})(\mathbf{I}+\bar{\mathbf{A}})^{-1}], \tag{147}$$

in the limit of large n, where $\frac{1}{n}\mathbb{E}\mathbf{H}^T\mathbf{H}=\mathbf{I}$ and $\frac{1}{n}\mathbb{E}\mathbf{N}^T\mathbf{N}=\mathbf{\Sigma}$. Furthermore, Moreau envelopes of Gaussian variables will concentrate on their expected values as the dimensions of the problems increases, see for example (Loureiro et al., 2021)[Supplement, lemma 5.]. As such the concentration of F to $\bar{F}$ is pointwise. Using similar arguments to lemma 2 we can bound $\mathbf{A},\bar{\mathbf{A}},\mathbf{B},\bar{\mathbf{B}}$ to convex sets. We can then find from the well known result that convergence of convex functions on convex sets implies uniform convergence, (Andersen and Gill, 1982)[appendix 2], to conclude that

$$\min_{\mathbf{B},\bar{\mathbf{A}}}\max_{\mathbf{A},\bar{\mathbf{B}}} F \xrightarrow[n,m\to\infty]{P} \min_{\mathbf{B},\bar{\mathbf{A}}}\max_{\mathbf{A},\bar{\mathbf{B}}} \bar{F}. \tag{148}$$

□

As such we obtain the following optimization:

$$\min_{\mathbf{B},\bar{\mathbf{A}}}\max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2K}\text{Tr}[\bar{\mathbf{A}}\mathbf{A}^T\mathbf{A}-\bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}]+\frac{1}{mK}\mathbb{E}\mathcal{M}_{\bar{\mathbf{B}}\ R}\left(\mathbf{\Theta}^*-\sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right) -\frac{m}{2nK}\text{Tr}[\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}]+\frac{1}{2K}\text{Tr}[(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^T(\mathbf{I}+\bar{\mathbf{A}})^{-1}]. \tag{149}$$

We now wish to solve over $\bar{\mathbf{A}}$. To accomplish this we need to interchange the order of the min and max, this can be justified, principally on the convexity/concavity of the problem and the fact that the problem can be reduced to compact convex sets. We do not prove this here, but refer to (Loureiro et al., 2021)[supplement section B.4] where a similar argument is used in their simplification of the the CGMT alternative. We now solve over $\bar{\mathbf{A}}$ to find:

$$\bar{\mathbf{A}}=\mathbf{A}^{-1}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}-\mathbf{I}. \tag{150}$$

Substituting we obtain:

$$\min_{\mathbf{B}}\max_{\mathbf{A},\bar{\mathbf{B}}} -\frac{1}{2K}\text{Tr}[\mathbf{A}^T\mathbf{A}]-\frac{1}{2K}\text{Tr}[\bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}]+\frac{1}{mK}\mathbb{E}\mathcal{M}_{\bar{\mathbf{B}}\ R}\left(\mathbf{\Theta}^*-\sqrt{\frac{m}{n}}\mathbf{F}^T\mathbf{A}\bar{\mathbf{B}}^{-1}\right) -\frac{m}{2nK}\text{Tr}[\mathbf{A}\bar{\mathbf{B}}^{-1}\mathbf{A}]+\frac{1}{K}\text{Tr}[\mathbf{A}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}]. \tag{151}$$

We can now specialize to the case of $R=\frac{1}{2}\text{Tr}[\mathbf{\Theta}\mathbf{\Lambda}\mathbf{\Theta}^T]$ where $\mathbf{\Lambda}\in\mathbb{R}^{K\times K}$. We return to equation 140 and substitute in this value

$$\min_{\mathbf{E},\mathbf{B}}\max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2mK}\text{Tr}[\mathbf{\Theta}^*\mathbf{\Lambda}\mathbf{\Theta}^{*T}]-\frac{1}{2K}\text{Tr}[\mathbf{A}^T\mathbf{A}]-\frac{1}{2K}\text{Tr}[\bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}]+\frac{1}{K\sqrt{mn}}\text{Tr}[\mathbf{A}\mathbf{F}\mathbf{E}]+\frac{1}{2mK}\text{Tr}[\bar{\mathbf{B}}\mathbf{E}^T\mathbf{E}] +\frac{1}{2mK}\text{Tr}[\mathbf{E}\mathbf{\Lambda}\mathbf{E}^T]+\frac{1}{mK}\text{Tr}[\mathbf{E}\mathbf{\Lambda}\mathbf{\Theta}^*]+\frac{1}{K}\text{Tr}[\mathbf{A}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}]. \tag{152}$$

we can now solve over $\mathbf{E}$ to find

$$\mathbf{E}=\left(\bar{\mathbf{B}}+\mathbf{\Lambda}\right)^{-1}\left[\mathbf{\Lambda}\mathbf{\Theta}^*+\sqrt{\frac{m}{n}}\mathbf{A}\mathbf{F}\right], \tag{153}$$

and substituting in this value we obtain:

$$\min_{\mathbf{B}}\max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2mK}\text{Tr}[\mathbf{\Theta}^*\mathbf{\Lambda}\mathbf{\Theta}^{*T}]-\frac{1}{2K}\text{Tr}[\mathbf{A}^T\mathbf{A}]-\frac{1}{2K}\text{Tr}[\bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}]+\frac{1}{K}\text{Tr}[\mathbf{A}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}] -\frac{1}{2mK}\text{Tr}[\left(\mathbf{\Lambda}\mathbf{\Theta}^*+\sqrt{\frac{m}{n}}\mathbf{A}\mathbf{F}\right)\left(\bar{\mathbf{B}}+\mathbf{\Lambda}\right)^{-1}\left(\mathbf{\Lambda}\mathbf{\Theta}^*+\sqrt{\frac{m}{n}}\mathbf{A}\mathbf{F}\right)]. \tag{154}$$

Using the same arguments as lemma 3, this concentrates to:

$$\min_{\mathbf{B}} \max_{\mathbf{A},\bar{\mathbf{B}}} \frac{1}{2mK}\mathrm{Tr}[\mathbf{\Theta}^*\mathbf{\Lambda}\mathbf{\Theta}^{*T}] - \frac{1}{2K}\mathrm{Tr}[\mathbf{A}^T\mathbf{A}] - \frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{B}}\mathbf{B}^T\mathbf{B}] + \frac{1}{K}\mathrm{Tr}[\mathbf{A}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}] \\ -\frac{1}{2K}\mathrm{Tr}[\left(\frac{1}{m}\mathbf{\Lambda}\mathbf{\Theta}^{*T}\mathbf{\Theta}^*\mathbf{\Lambda} + \frac{m}{n}\mathbf{A}\mathbf{A}\right)(\bar{\mathbf{B}}+\mathbf{\Lambda})^{-1}]. \tag{155}$$

We can now solve over $\bar{\mathbf{B}}$ to fin that

$$\bar{\mathbf{B}} = \mathbf{B}^{-1}\left(\frac{1}{m}\mathbf{\Lambda}\mathbf{\Theta}^{*T}\mathbf{\Theta}^*\mathbf{\Lambda} + \frac{m}{n}\mathbf{A}\mathbf{A}\right)^{1/2} - \mathbf{\Lambda}, \tag{156}$$

from which we obtain:

$$\min_{\mathbf{B}} \max_{\mathbf{A}} \frac{1}{2mK}\mathrm{Tr}[\mathbf{\Theta}^*\mathbf{\Lambda}\mathbf{\Theta}^{*T}] + \frac{1}{2K}\mathrm{Tr}[\mathbf{\Lambda}\mathbf{B}^T\mathbf{B}] - \frac{1}{2K}\mathrm{Tr}[\mathbf{A}^T\mathbf{A}] + \frac{1}{K}\mathrm{Tr}[\mathbf{A}(\mathbf{B}\mathbf{B}+\mathbf{\Sigma})^{1/2}] \\ -\frac{1}{K}\mathrm{Tr}[\mathbf{B}\left(\frac{1}{m}\mathbf{\Lambda}\mathbf{\Theta}^{*T}\mathbf{\Theta}^*\mathbf{\Lambda} + \frac{m}{n}\mathbf{A}\mathbf{A}\right)^{1/2}]. \tag{157}$$

F.2 Complex Regression

We recall the optimization of interest:

$$\min_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\mathbb{R}^m} \max_{\mathbf{z}_1,\mathbf{z}_2\in\mathbb{R}^n} \frac{1}{4n\sqrt{m}} \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{G} & -\mathbf{H}\\ \mathbf{H} & \mathbf{G}\end{bmatrix} \begin{bmatrix}\boldsymbol{\theta}_1\\ \boldsymbol{\theta}_2\end{bmatrix} - \mathbf{z}^T\mathbf{y} - \frac{1}{4n}\|\mathbf{z}\| + \frac{1}{2m}R(\boldsymbol{\theta}). \tag{158}$$

We recall that we choose the following model for the labels $\mathbf{y} = (\mathbf{G}+i\mathbf{H})\boldsymbol{\theta}^* + \boldsymbol{\nu}$, where $\boldsymbol{\nu} = \boldsymbol{\nu}_1 + i\boldsymbol{\nu}_2$ is zero mean complex noise with variances $\sigma^2_{\boldsymbol{\nu}_1}, \sigma^2_{\boldsymbol{\nu}_2}$. We therefore obtain the following model:

$$\min_{\mathbf{e}_1,\mathbf{e}_2\in\mathbb{R}^m} \max_{\mathbf{z}_1,\mathbf{z}_2\in\mathbb{R}^n} \frac{1}{4n\sqrt{m}} \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{G} & -\mathbf{H}\\ \mathbf{H} & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix} - \mathbf{z}^T\boldsymbol{\nu} - \frac{1}{4n}\|\mathbf{z}\| + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*), \tag{159}$$

where $\mathbf{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. In the following lemma we show that the problem can be restricted to compact and convex sets:

Lemma 4. *Let $\hat{\mathbf{e}}, \hat{\mathbf{z}}$ be the optimal values of (159). Then there exists positive constants $C_{\mathbf{e}}$ and $C_{\mathbf{z}}$ such that:*

$$\Pr\left(\|\hat{\mathbf{e}}\|_2 \le C_{\mathbf{e}}\sqrt{m}\right) \xrightarrow[m\to\infty]{P} 1, \qquad \Pr\left(\|\hat{\mathbf{z}}\|_2 \le C_{\mathbf{z}}\sqrt{n}\right) \xrightarrow[n\to\infty]{P} 1. \tag{160}$$

Proof. The proof is very similar to the proof of lemma 2, we do not reproduce it here. □

WE can therefore defin the sets $\mathcal{S}_{\mathbf{e}} = \{\,\mathbf{e}\in\mathbb{R}^{2m} \mid \|\mathbf{e}\| \le C\sqrt{m}\,\}$ and $\mathcal{S}_{\mathbf{z}} = \{\,\mathbf{z}\in\mathbb{R}^{2n} \mid \|\mathbf{z}\|_2 \le C\sqrt{n}\,\}$ for some constant C. Our optimization is therefore of the form:

$$\min_{\mathbf{e}\in\mathcal{S}_{\mathbf{e}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \frac{1}{4n\sqrt{m}} \begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{G} & -\mathbf{H}\\ \mathbf{H} & \mathbf{G}\end{bmatrix} \begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\end{bmatrix} - \mathbf{z}^T\boldsymbol{\nu} - \frac{1}{4n}\|\mathbf{z}\| + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{161}$$

The objective is now in a form where we can apply a special case of our CGMT extension, corollary 2, we obtain the following optimization:

$$\min_{\mathbf{e}\in\mathcal{S}_{\mathbf{e}}} \max_{\mathbf{z}\in\mathcal{S}_{\mathbf{z}}} \frac{1}{2n\sqrt{2m}}\|\mathbf{e}\|_2\,\mathbf{z}^T\mathbf{g} + \frac{1}{2n\sqrt{2m}}\|\mathbf{z}\|\,\mathbf{h}^T\mathbf{e} - \frac{1}{2n}\mathbf{z}^T\boldsymbol{\nu} - \frac{1}{4n}\|\mathbf{z}\|^2 + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{162}$$

we now define $\beta = \frac{1}{\sqrt{2n}}\|\mathbf{z}\|_2$ and solve over $\mathbf{z}$:

$$\min_{\mathbf{e}\in\mathcal{S}_{\mathbf{e}}} \max_{\beta\ge 0} \frac{\beta}{\sqrt{4nm}}\mathbf{h}^T\mathbf{e} + \frac{\beta}{\sqrt{2n}}\left\|\frac{1}{\sqrt{2m}}\|\mathbf{e}\|\,\mathbf{g} - \boldsymbol{\nu}\right\|_2 - \frac{\beta^2}{2} + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{163}$$

Using similar arguments as in lemma 3 we can show that in the large n, m regime the 2-norm will concentrate on its expected value. As such we consider the following optimization

$$\min_{\mathbf{e}\in\mathcal{S}_\mathbf{e}} \max_{\beta\geq 0} \frac{\beta}{\sqrt{4nm}}\mathbf{h}^T\mathbf{e} + \beta\sqrt{\frac{1}{2m}\left\|\mathbf{e}\right\|^2 + \sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2} - \frac{\beta^2}{2} + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{164}$$

We then introduce q using the square root trick:

$$\min_{\mathbf{e}\in\mathcal{S}_\mathbf{e}, q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} + \frac{\beta}{\sqrt{4nm}}\mathbf{h}^T\mathbf{e} + \frac{\beta}{4qm}\left\|\mathbf{e}\right\|^2 + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{165}$$

We can now complete the square over $\mathbf{e}$

$$\min_{\mathbf{e}\in\mathcal{S}_\mathbf{e}, q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} + \frac{\beta}{4qm}\left\|\mathbf{e} + \frac{q\sqrt{m}}{\sqrt{n}}\mathbf{h}\right\|^2 - \frac{\beta q m}{4n^2}\left\|\mathbf{h}\right\|^2 + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{1}{2m}R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{166}$$

We then reintroduce $\boldsymbol{\theta} = \mathbf{e} + \boldsymbol{\theta}^*$ and note the Moreau envelope over $\mathbf{e}$

$$\min_{q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} - \frac{\beta q m}{4n^2}\left\|\mathbf{h}\right\|^2 + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{1}{2m}\mathcal{M}_{\frac{q}{\beta}\ R}\left(\boldsymbol{\theta}^* - \frac{q\sqrt{m}}{\sqrt{n}}\mathbf{h}\right). \tag{167}$$

Finally using similar concentration arguments as lemma 3, we obtain the following optimization:

$$\min_{q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} - \frac{\beta q m}{2n} + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{1}{2m}\mathbb{E}\mathcal{M}_{\frac{q}{\beta}\ R}\left(\boldsymbol{\theta}^* - \frac{q\sqrt{m}}{\sqrt{n}}\mathbf{h}\right). \tag{168}$$

Furthermore, when specializing to the case that $R(\boldsymbol{\theta}) = \frac{\lambda}{2}\left\|\boldsymbol{\theta}\right\|^2$, we can find the following. Returning to equation (165), we find

$$\min_{\mathbf{e}\in\mathcal{S}_\mathbf{e}, q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} + \frac{\beta}{\sqrt{4nm}}\mathbf{h}^T\mathbf{e} + \frac{\beta}{4qm}\left\|\mathbf{e}\right\|^2 + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{\lambda}{4m}\left\|\mathbf{e}\right\|^2 + \frac{\lambda}{2m}\mathbf{e}^T\boldsymbol{\theta}^* + \frac{\lambda}{4m}\left\|\boldsymbol{\theta}^*\right\|^2. \tag{169}$$

Solving over $\mathbf{e}$ we find:

$$\mathbf{e} = -\frac{q}{\beta+\lambda q}\left[\lambda\boldsymbol{\theta}^* + \frac{\beta\sqrt{m}}{\sqrt{n}}\mathbf{h}\right]. \tag{170}$$

Substituting and taking the concentration we finally end up with:

$$\min_{q\geq 0} \max_{\beta\geq 0} \frac{\beta q}{2} + \frac{\beta(\sigma_{\boldsymbol{\nu}_1}^2 + \sigma_{\boldsymbol{\nu}_2}^2)}{2q} - \frac{\beta^2}{2} + \frac{\beta\lambda}{4m(\beta+\lambda q)}\left\|\boldsymbol{\theta}^*\right\|^2 + \frac{q\beta^2 m}{2n(\beta+q\lambda)}. \tag{171}$$

F.3 General Applications

For the application to the case of convolutional regression we can note that the generic version of the problem that do not have a special case as given by corollaries 1,2, can be expressed as:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^M} \frac{1}{2N}\left\|\frac{1}{\sqrt{M}}\bar{\mathbf{X}}\boldsymbol{\theta} - \mathbf{y}\right\|^2 + R(\boldsymbol{\theta}), \tag{172}$$

where $\bar{\mathbf{X}} \in \mathbb{R}^{N\times M}$ and $\mathbf{y}\in\mathbb{R}^N$. In the case of convolutional regression $N = nD_1D_2$ and $M = k_1k_2$, and $\bar{\mathbf{X}} \in \mathbb{R}^{nD_1D_2\times k_1k_2}$ is defined element wise by:

$$(\bar{\mathbf{X}})_{\alpha D_1D_2+\beta D_2+\gamma,\eta k_2+\epsilon} = \sum_{a=1}^{n}\sum_{b=1}^{d_1}\sum_{c=1}^{d_2}(X_a)_{bc}\delta_{a,\alpha}\delta_{\beta+\eta,b}\delta_{\gamma+\epsilon,c}, \tag{173}$$

where each $\mathbf{X}_i \in \mathbb{R}^{d_1 \times d_2}$ for $i \in [n]$ had i.i.d. standard Gaussian elements. We consider the generic setup of equation (172) first before specializing. We can first note that we can rexpress this optimization as a min-max problem:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^M} \max_{\mathbf{z} \in \mathbb{R}^N} \frac{1}{N\sqrt{M}} \mathbf{z}^T \bar{\mathbf{X}} \boldsymbol{\theta} - \frac{1}{N} \mathbf{z}^T \mathbf{y} - \frac{1}{2N} \|\mathbf{z}\|_2^2 + R(\boldsymbol{\theta}). \tag{174}$$

We now assume that $\bar{\mathbf{X}}$ is a GMS of order $K \in \mathbb{N}$ and takes the form $\bar{\mathbf{X}} = \sum_{k=1}^K \mathbf{A}_k^T \mathbf{X} \mathbf{B}_k$ for some $\mathbf{X} \in \mathbb{R}^{\tilde{N} \times \tilde{M}}$ and each $\mathbf{A}_k \in \mathbb{R}^{\tilde{N} \times N}$ and $\mathbf{B}_k \in \mathbb{R}^{\tilde{M} \times M}$. We will further assume that $\mathbf{y} = \frac{1}{\sqrt{M}} \bar{\mathbf{X}} \boldsymbol{\theta}^* + \boldsymbol{\nu}$, for some true parameter $\boldsymbol{\theta}^*$ and some zero mean noise $\boldsymbol{\nu} \in \mathbb{R}^N$ with covariance $\boldsymbol{\Sigma}$. We therefore define the error vector $\mathbf{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$. In the lemma below we show that this optimization can be restricted to compact sets:

Lemma 5. *Let $\hat{\mathbf{e}}, \hat{\mathbf{z}}$ be the optimal values of* (174)*. Then there exists positive constants $C_{\mathbf{e}}$ and $C_{\mathbf{z}}$ such that:*

$$\Pr\left(\|\hat{\mathbf{e}}\|_2 \leq C_{\mathbf{e}} \sqrt{M} \right) \xrightarrow[M \to \infty]{P} 1, \qquad \Pr\left(\|\hat{\mathbf{z}}\|_2 \leq C_{\mathbf{z}} \sqrt{N} \right) \xrightarrow[N \to \infty]{P} 1. \tag{175}$$

Proof. The proof is very similar to the proof of lemma 2. □

As such we can construct two sets $\mathcal{S}_{\mathbf{e}} = \{ \mathbf{e} \in \mathbb{R}^M \mid \|\mathbf{e}\|_2 \leq C\sqrt{M} \}$ and $\mathcal{S}_{\mathbf{z}} = \{ \mathbf{z} \in \mathbb{R}^N \mid \|\mathbf{z}\|_2 \leq C\sqrt{n} \}$. Our optimization takes the following form:

$$\min_{\mathbf{e} \in \mathcal{S}_{\mathbf{e}}} \max_{\mathbf{z} \in \mathcal{S}_{\mathbf{z}}} \frac{1}{N\sqrt{M}} \mathbf{z}^T \bar{\mathbf{X}} \mathbf{e} - \frac{1}{N} \mathbf{z}^T \boldsymbol{\nu} - \frac{1}{2N} \|\mathbf{z}\|_2^2 + R(\boldsymbol{\theta}). \tag{176}$$

We can now apply our novel CGMT extention to obtain the following alernative optimization

$$\min_{\mathbf{e} \in \mathcal{S}_{\mathbf{e}}} \max_{\mathbf{z} \in \mathcal{S}_{\mathbf{z}}} \frac{1}{N\sqrt{M}} \sum_{k=1}^K \mathbf{f}_k^T \mathbf{B}_k \mathbf{e} + \frac{1}{N\sqrt{M}} \sum_{k=1}^K \mathbf{h}_k^T \mathbf{A}_k \mathbf{z} - \frac{1}{N} \mathbf{z}^T \boldsymbol{\nu} - \frac{1}{2N} \|\mathbf{z}\|_2^2 + R(\boldsymbol{\theta}). \tag{177}$$

Next, we recall that $\mathbb{E}[\mathbf{f}_k \mathbf{f}_{k'}] = \mathbf{z}^T \mathbf{A}_k^T \mathbf{A}_{k'} \mathbf{z}$ and $\mathbb{E}[\mathbf{h}_k \mathbf{h}_{k'}] = \mathbf{e}^T \mathbf{B}_k^T \mathbf{B}_{k'} \mathbf{e}$. We now define $P_{k,k'} = \frac{1}{N} \mathbf{z}^T \mathbf{A}_k^T \mathbf{A}_{k'} \mathbf{z}$ and $Q_{k,k'} = \frac{1}{M} \mathbf{e}^T \mathbf{B}_k^T \mathbf{B}_{k'} \mathbf{e}$, where $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{K \times K}$. We note that these are PSD matrices. We reintroduce these constraints with PSD lagrange multiplier $\bar{\mathbf{P}}$ and $\bar{\mathbf{Q}}$ respectively:

$$\begin{aligned} \min_{\mathbf{Q}, \bar{\mathbf{P}} \in \mathbb{R}^{K \times K}} \max_{\mathbf{P}, \bar{\mathbf{Q}} \in \mathbb{R}^{K \times K}} \min_{\mathbf{e} \in \mathcal{S}_{\mathbf{e}}} \max_{\mathbf{z} \in \mathcal{S}_{\mathbf{z}}} & \frac{1}{\sqrt{NM}} \sum_{k=1}^K \mathbf{f}_k^T \mathbf{B}_k \mathbf{e} + \frac{1}{N} \sum_{k=1}^K \mathbf{h}_k^T \mathbf{A}_k \mathbf{z} - \frac{1}{N} \mathbf{z}^T \boldsymbol{\nu} - \frac{1}{2N} \|\mathbf{z}\|_2^2 + R(\boldsymbol{\theta}) \\ & + \frac{1}{2K} \sum_{k,k'}^K \bar{P}_{k,k'} \left(P_{k,k'} - \frac{1}{N} \mathbf{z}^T \mathbf{A}_k^T \mathbf{A}_{k'} \mathbf{z} \right) + \frac{1}{2K} \sum_{k,k'}^K \bar{Q}_{k,k'} \left(\frac{1}{M} \mathbf{e}^T \mathbf{B}_k^T \mathbf{B}_{k'} \mathbf{e} - Q_{k,k'} \right). \end{aligned} \tag{178}$$

We now note that $\mathbb{E}[\mathbf{f}_k \mathbf{f}_{k'}] = P_{k,k'}$ and $\mathbb{E}[\mathbf{h}_k \mathbf{h}_{k'}] = Q_{k,k'}$. We can now solve over $\mathbf{z}$, to find that:

$$\mathbf{z} = \left(\mathbf{I} + \frac{1}{K} \sum_{k,k'}^K \bar{P}_{k,k'} \mathbf{A}_k^T \mathbf{A}_{k'} \right)^{-1} \left(\sum_{k=1}^K \mathbf{A}_k^T \mathbf{h}_k - \boldsymbol{\nu} \right). \tag{179}$$

Substituting in this value, we obtain:

$$\begin{aligned} \min_{\mathbf{Q}, \bar{\mathbf{P}} \in \mathbb{R}^{K \times K}} \max_{\mathbf{P}, \bar{\mathbf{Q}} \in \mathbb{R}^{K \times K}} \min_{\mathbf{e} \in \mathcal{S}_{\mathbf{e}}} & \frac{1}{\sqrt{NM}} \sum_{k=1}^K \mathbf{f}_k^T \mathbf{B}_k \mathbf{e} + R(\boldsymbol{\theta}) + \frac{1}{2K} \mathrm{Tr}[\bar{\mathbf{P}} \mathbf{P} - \bar{\mathbf{Q}} \mathbf{Q}] + \frac{1}{2KM} \sum_{k,k'}^K \bar{Q}_{k,k'} \mathbf{e}^T \mathbf{B}_k^T \mathbf{B}_{k'} \mathbf{e} \\ & + \frac{1}{2N} \left(\sum_{k=1}^K \mathbf{A}_k^T \mathbf{h}_k - \boldsymbol{\nu} \right)^T \left(\mathbf{I} + \frac{1}{K} \sum_{k,k'}^K \bar{P}_{k,k'} \mathbf{A}_k^T \mathbf{A}_{k'} \right)^{-1} \left(\sum_{k=1}^K \mathbf{A}_k^T \mathbf{h}_k - \boldsymbol{\nu} \right). \end{aligned} \tag{180}$$

We can now complete the square over $\mathbf{e}$ to find that:

$$\begin{aligned}
&\min_{\mathbf{Q},\bar{\mathbf{P}}\in\mathbb{R}^{K\times K}}\max_{\mathbf{P},\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K}}\min_{\mathbf{e}\in\mathcal{S}_{\mathbf{e}}} R(\boldsymbol{\theta})+\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{P}}\mathbf{P}-\bar{\mathbf{Q}}\mathbf{Q}]\\
&\frac{1}{2M}\left(\mathbf{e}+\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\right)\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\\
&\qquad\times\left(\mathbf{e}+\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\right)\\
&\qquad-\frac{1}{2N}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)^T\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\\
&\quad+\frac{1}{2N}\left(\sum_{k=1}^K\mathbf{A}_k^T\mathbf{h}_k-\boldsymbol{\nu}\right)^T\left(\mathbf{I}+\frac{1}{K}\sum_{k,k'}\bar{P}_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{A}_k^T\mathbf{h}_k-\boldsymbol{\nu}\right).
\end{aligned}\tag{181}$$

We can recognize the Matrix Moreau envelope over $\mathbf{e}$, simplifying we obtain:

$$\begin{aligned}
&\min_{\mathbf{Q},\bar{\mathbf{P}}\in\mathbb{R}^{K\times K}}\max_{\mathbf{P},\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K}}\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{P}}\mathbf{P}-\bar{\mathbf{Q}}\mathbf{Q}]\\
&\frac{1}{M}\mathcal{M}_{\left(\frac{1}{K}\sum_{k,k'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\ R(\cdot+\boldsymbol{\theta}^*)}\left(-\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\right)\\
&\qquad-\frac{1}{2N}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)^T\left(\frac{1}{K}\sum_{k,k'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\\
&\quad+\frac{1}{2N}\left(\sum_{k=1}^K\mathbf{A}_k^T\mathbf{h}_k-\boldsymbol{\nu}\right)^T\left(\mathbf{I}+\frac{1}{K}\sum_{k,k'}^K\bar{P}_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{A}_k^T\mathbf{h}_k-\boldsymbol{\nu}\right).
\end{aligned}\tag{182}$$

Using similar concentration arguments as used in lemma 3, we can note the concentration of the Moreau envelope and the quadratic form, we therefore consider instead the following optimization:

$$\begin{aligned}
&\min_{\mathbf{Q},\bar{\mathbf{P}}\in\mathbb{R}^{K\times K}}\max_{\mathbf{P},\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K}}\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{P}}\mathbf{P}-\bar{\mathbf{Q}}\mathbf{Q}]\\
&\frac{1}{M}\mathcal{M}_{\left(\frac{1}{K}\sum_{k,k'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\ R(\cdot+\boldsymbol{\theta}^*)}\left(-\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^K\mathbf{B}_k^T\mathbf{f}_k\right)\right)\\
&\qquad-\frac{1}{2N}\mathrm{Tr}\left[\left(\sum_{k,k'}P_{k,k'}\mathbf{B}_k^T\mathbf{B}k'\right)\left(\frac{1}{K}\sum_{l,l'}^K\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\right]\\
&\quad+\frac{1}{2N}\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\left(\mathbf{I}+\frac{1}{K}\sum_{k,k'}\bar{P}_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{-1}\right].
\end{aligned}\tag{183}$$

To simplify this further we let $\mathbf{C}=\mathbf{I}+\frac{1}{K}\sum_{k,k'}\bar{P}_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}$ and reintroduce this constraint with a Lagrange

multiplier $\mathbf{D}$

$$\min_{\mathbf{Q},\bar{\mathbf{P}}\in\mathbb{R}^{K\times K},\mathbf{C}\in\mathbb{R}^{N\times N}}\max_{\mathbf{P},\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K},\mathbf{D}\in\mathbb{R}^{N\times N}}\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{P}}\mathbf{P}-\bar{\mathbf{Q}}\mathbf{Q}]$$
$$\frac{1}{M}\mathcal{M}_{\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\ R(\cdot+\boldsymbol{\theta}^*)}\left(-\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)\right)$$
$$-\frac{1}{2N}\mathrm{Tr}\left[\left(\sum_{k,k'}P_{k,k'}\mathbf{B}_k^T\mathbf{B}k'\right)\left(\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\right]$$
$$+\frac{1}{2N}\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{C}^{-1}\right]+\frac{1}{2N}\mathrm{Tr}\mathbf{D}\mathbf{C}-\frac{1}{2N}\mathrm{Tr}\mathbf{D}-\frac{1}{2KN}\sum_{k,k'}\bar{P}_{k,k'}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}. \tag{184}$$

We can then solve over $\bar{\mathbf{P}}$, to find that:

$$P_{k,k'}=\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}. \tag{185}$$

Substituting we can find that:

$$\min_{\mathbf{Q}\in\mathbb{R}^{K\times K},\mathbf{C}\in\mathbb{R}^{N\times N}}\max_{\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K},\mathbf{D}\in\mathbb{R}^{N\times N}}-\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{Q}}\mathbf{Q}]$$
$$\frac{1}{M}\mathcal{M}_{\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\ R(\cdot+\boldsymbol{\theta}^*)}\left(-\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)\right)$$
$$-\frac{1}{2N}\mathrm{Tr}\left[\left(\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}k'\right)\left(\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\right]$$
$$+\frac{1}{2N}\mathrm{Tr}\left[\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{C}^{-1}\right]+\frac{1}{2N}\mathrm{Tr}\mathbf{D}\mathbf{C}-\frac{1}{2N}\mathrm{Tr}\mathbf{D}. \tag{186}$$

We can now solve over $\mathbf{C}$ to find that:

$$\mathbf{C}=\mathbf{D}^{-1/2}\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}. \tag{187}$$

Which after substitution, we obtain:

$$\min_{\mathbf{Q}\in\mathbb{R}^{K\times K}}\max_{\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K},\mathbf{D}\in\mathbb{R}^{N\times N}}-\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{Q}}\mathbf{Q}]-\frac{1}{2N}\mathrm{Tr}\mathbf{D}$$
$$\frac{1}{M}\mathcal{M}_{\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\ R(\cdot+\boldsymbol{\theta}^*)}\left(-\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)\right)$$
$$-\frac{1}{2N}\mathrm{Tr}\left[\left(\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}k'\right)\left(\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\right]$$
$$+\frac{1}{N}\mathrm{Tr}\mathbf{D}^{1/2}\left[\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]. \tag{188}$$

For Convolution, we also specifically consider the case where the regression is a square function. We will choose $R(\boldsymbol{\theta}) = \frac{1}{2M}\boldsymbol{\theta}^T\boldsymbol{\Lambda}\boldsymbol{\theta}$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{M\times M}$, and specialize the form of $\boldsymbol{\Lambda}$ in each of the cases discussed below. The Moreau envelope for this choice takes the form:

$$\min_{\mathbf{e}\in\mathcal{S}_{\mathbf{e}}} \frac{1}{2M}\left(\mathbf{e}+\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)\right)\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)$$
$$\times\left(\mathbf{e}+\sqrt{\frac{M}{N}}\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)\right)+\frac{1}{2M}(\mathbf{e}+\boldsymbol{\theta}^*)^T\boldsymbol{\Lambda}(\mathbf{e}+\boldsymbol{\theta}^*). \tag{189}$$

Which has the optimal solution:

$$\mathbf{e} = -\left(\boldsymbol{\Lambda}+\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\Lambda\boldsymbol{\theta}^*+\sqrt{\frac{M}{N}}\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right). \tag{190}$$

and optimal value:

$$\frac{1}{2M}\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda}\boldsymbol{\theta}^*+\frac{1}{2N}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)^T\left(\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)$$
$$-\frac{1}{2M}\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*+\sqrt{\frac{M}{N}}\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right)^T\left(\boldsymbol{\Lambda}+\frac{1}{K}\sum_{k,k'}\bar{Q}_{k,k'}\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{-1}\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*+\sqrt{\frac{M}{N}}\sum_{k=1}^{K}\mathbf{B}_k^T\mathbf{f}_k\right). \tag{191}$$

By arguments similar to those used in lemma 3, this value will concentrate on:

$$\frac{1}{2M}\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda}\boldsymbol{\theta}^*-\frac{1}{2N}\mathrm{Tr}\left[\left(\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}k'\right)\left(\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{l,l'}\mathbf{B}_l^T\mathbf{B}_{l'}\right)^{-1}\right]$$
$$-\frac{1}{2M}\mathrm{Tr}\left[\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda}+\frac{M}{N}\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}_{k'}\right)\left(\boldsymbol{\Lambda}+\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{l,l'}\mathbf{B}_l^T\mathbf{B}_{l'}\right)^{-1}\right]. \tag{192}$$

Substituting in this value we obtain:

$$\min_{\mathbf{Q}\in\mathbb{R}^{K\times K}}\max_{\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K},\mathbf{D}\in\mathbb{R}^{N\times N}}\frac{1}{2M}\boldsymbol{\theta}^*\boldsymbol{\Lambda}\boldsymbol{\theta}^*-\frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{Q}}\mathbf{Q}]-\frac{1}{2N}\mathrm{Tr}\mathbf{D}$$
$$-\frac{1}{2M}\mathrm{Tr}\left[\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda}+\frac{M}{N}\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}_{k'}\right)\left(\boldsymbol{\Lambda}+\frac{1}{K}\sum_{l,l'}^{K}\bar{Q}_{l,l'}\mathbf{B}_l^T\mathbf{B}_{l'}\right)^{-1}\right]$$
$$+\frac{1}{N}\mathrm{Tr}\mathbf{D}^{1/2}\left[\left(\boldsymbol{\Sigma}+\sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]. \tag{193}$$

We next define $\mathbf{E} = \mathbf{\Lambda} + \frac{1}{K}\sum_{l,l'}^{K} \bar{Q}_{l,l'}\mathbf{B}_l^T\mathbf{B}_{l'}$ and reintroduce this constraint with $\mathbf{F}$

$$\min_{\mathbf{Q}\in\mathbb{R}^{K\times K},\mathbf{F}\in\mathbb{R}^{M\times M}} \max_{\bar{\mathbf{Q}}\in\mathbb{R}^{K\times K},\mathbf{D}\in\mathbb{R}^{N\times N},\mathbf{E}\in\mathbb{R}^{M\times M}} \frac{1}{2M}\boldsymbol{\theta}^*\mathbf{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2K}\mathrm{Tr}[\bar{\mathbf{Q}}\mathbf{Q}] - \frac{1}{2N}\mathrm{Tr}\mathbf{D} - \frac{1}{2M}\mathrm{Tr}\mathbf{F}\mathbf{E} + \frac{1}{2M}\mathrm{Tr}\mathbf{F}\mathbf{\Lambda}$$
$$-\frac{1}{2M}\mathrm{Tr}\left[\left(\mathbf{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\mathbf{\Lambda} + \frac{M}{N}\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}_{k'}\right)\mathbf{E}^{-1}\right] + \frac{1}{2MK}\sum_{k,k'}^{K}\bar{Q}_{k,k'}\mathrm{Tr}\mathbf{F}\mathbf{B}_k^T\mathbf{B}_{k'}$$
$$+\frac{1}{N}\mathrm{Tr}\mathbf{D}^{1/2}\left[\left(\mathbf{\Sigma} + \sum_{k,k'}Q_{k,k'}\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]. \quad (194)$$

We can now solve over $\bar{\mathbf{Q}}$ to obtain:

$$Q_{k,k'} = \frac{1}{M}\mathrm{Tr}\mathbf{F}\mathbf{B}_k^T\mathbf{B}_{k'}. \quad (195)$$

Substituting we obtain:

$$\min_{\mathbf{F}\in\mathbb{R}^{M\times M}} \max_{\mathbf{D}\in\mathbb{R}^{N\times N},\mathbf{E}\in\mathbb{R}^{M\times M}} \frac{1}{2M}\boldsymbol{\theta}^*\mathbf{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2N}\mathrm{Tr}\mathbf{D} - \frac{1}{2M}\mathrm{Tr}\mathbf{F}\mathbf{E} + \frac{1}{2M}\mathrm{Tr}\mathbf{F}\mathbf{\Lambda}$$
$$-\frac{1}{2M}\mathrm{Tr}\left[\left(\mathbf{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\mathbf{\Lambda} + \frac{M}{N}\sum_{k,k'}\left(\frac{1}{N}\mathrm{Tr}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}_{k'}\right)\mathbf{E}^{-1}\right]$$
$$+\frac{1}{N}\mathrm{Tr}\mathbf{D}^{1/2}\left[\left(\mathbf{\Sigma} + \sum_{k,k'}\left(\frac{1}{M}\mathrm{Tr}\mathbf{F}\mathbf{B}_k^T\mathbf{B}_{k'}\right)\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]. \quad (196)$$

Finally we can solve over $\mathbf{E}$ to obtain:

$$\mathbf{E} = \mathbf{F}^{-1/2}\left(\mathbf{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^*\mathbf{\Lambda} + \frac{M}{N}\sum_{k,k'}\left(\frac{1}{N}\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}\right)\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{1/2}. \quad (197)$$

Which after substitution gives us:

$$\min_{\mathbf{F}\in\mathbb{R}^{M\times M}} \max_{\mathbf{D}\in\mathbb{R}^{N\times N}} \frac{1}{2M}\boldsymbol{\theta}^*\mathbf{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2N}\mathrm{Tr}\mathbf{D} + \frac{1}{2M}\mathrm{Tr}\mathbf{F}\mathbf{\Lambda}$$
$$+\frac{1}{N}\mathrm{Tr}\left[\mathbf{D}^{1/2}\left(\mathbf{\Sigma} + \frac{1}{M}\sum_{k,k'}\mathrm{Tr}[\mathbf{F}\mathbf{B}_k^T\mathbf{B}_{k'}]\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]$$
$$-\frac{1}{M}\mathrm{Tr}\left[\mathbf{F}^{1/2}\left(\mathbf{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\mathbf{\Lambda} + \frac{M}{N^2}\sum_{k,k'}\mathrm{Tr}[\mathbf{D}\mathbf{A}_k^T\mathbf{A}_{k'}]\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{1/2}\right]. \quad (198)$$

Finally we introduce $\mathbf{S}, \mathbf{T}$ such that $\mathbf{S}^T\mathbf{S} = \mathbf{F}$ and $\mathbf{T}^T\mathbf{T} = \mathbf{D}$, so that our optimization can be expressed as:

$$\min_{\mathbf{S}\in\mathbb{R}^{M\times M}} \max_{\mathbf{T}\in\mathbb{R}^{N\times N}} \frac{1}{2M}\boldsymbol{\theta}^{*T}\mathbf{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2N}\mathrm{Tr}[\mathbf{T}^T\mathbf{T}] + \frac{1}{2M}\mathrm{Tr}[\mathbf{S}^T\mathbf{S}\mathbf{\Lambda}]$$
$$+\frac{1}{N}\mathrm{Tr}\left[\mathbf{T}\left(\mathbf{\Sigma} + \frac{1}{M}\sum_{k,k'}\mathrm{Tr}[\mathbf{S}^T\mathbf{S}\mathbf{B}_k^T\mathbf{B}_{k'}]\mathbf{A}_k^T\mathbf{A}_{k'}\right)^{1/2}\right]$$
$$-\frac{1}{M}\mathrm{Tr}\left[\mathbf{S}\left(\mathbf{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\mathbf{\Lambda} + \frac{M}{N^2}\sum_{k,k'}\mathrm{Tr}[\mathbf{T}^T\mathbf{T}\mathbf{A}_k^T\mathbf{A}_{k'}]\mathbf{B}_k^T\mathbf{B}_{k'}\right)^{1/2}\right]. \quad (199)$$

F.3.1 Convolutional Regression

In this we recall the following, we have $K = k_1 k_2$ classes, the value of $N = nD_1D_2$ and the value of $M = k_1k_2$. We also recall that $\mathbf{A}_{\omega,\nu} \in \mathbb{R}^{nD_1D_2 \times nd_1d_2}$ and $\mathbf{b}_{\omega,\nu} \in \mathbb{R}^{1 \times k_1k_2}$ are defined by:

$$(\mathbf{A}_{\omega,\nu})_{\alpha D_1D_2+\beta D_2+\gamma, ad_1d_2+bd_2+c} = \delta_{a,\alpha}\delta_{\beta+\omega,b}\delta_{\gamma+\nu,c},$$
$$(\mathbf{b}_{\omega,\nu})_{\eta k_2+\epsilon} = \delta_{\omega,\eta}\delta_{\epsilon,\nu}, \tag{200}$$

where $a, \alpha \in [n], \beta \in [D_1], \gamma \in [D_2], b \in [d_1], c \in [d_2], \eta, \omega \in [k_1], \epsilon, \nu \in [k_2]$. Will also for convenience introduce $\tilde{\mathbf{A}}_{\omega,\nu} \in \mathbb{R}^{D_1D_2 \times d_1d_2}$ defined element wise by $(\mathbf{A}_{\omega,\nu})_{\beta D_2+\gamma, bd_2+c} = \delta_{\beta+\omega,b}\delta_{\gamma+\nu,c}$, such that $\mathbf{A}_{\omega,\nu} = \tilde{\mathbf{A}}_{\omega,\nu} \otimes \mathbf{I}_n$. We can then note that our optimization takes the form:

$$\min_{\mathbf{S}\in\mathbb{R}^{k_1k_2\times k_1k_2}} \max_{\mathbf{T}\in\mathbb{R}^{nD_1D_2\times nD_1D_2}} \frac{1}{2k_1k_2}\boldsymbol{\theta}^*\boldsymbol{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2nD_1D_2}\mathrm{Tr}[\mathbf{T}^T\mathbf{T}] + \frac{1}{2k_1k_2}\mathrm{Tr}[\mathbf{S}^T\mathbf{S}\boldsymbol{\Lambda}]$$
$$+\frac{1}{nD_1D_2}\mathrm{Tr}\left[\mathbf{T}\left(\boldsymbol{\Sigma} + \frac{1}{k_1k_2}\sum_{k,k'}\mathrm{Tr}[\mathbf{S}^T\mathbf{S}\mathbf{b}_{\omega,\nu}\mathbf{b}^T_{\omega',\nu'}]\mathbf{A}^T_{\omega,\nu}\mathbf{A}_{\omega',\nu'}\right)^{1/2}\right]$$
$$-\frac{1}{k_1k_2}\mathrm{Tr}\left[\mathbf{S}\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda} + \frac{k_1k_2}{n^2D_1^2D_2^2}\sum_{\omega,\omega'}^{k_1}\sum_{\nu,\nu'}^{k_2}\mathrm{Tr}[\mathbf{T}^T\mathbf{T}\mathbf{A}^T_{\omega,\nu'}\mathbf{A}_{\omega',\nu'}]\mathbf{b}_{\omega,\nu}\mathbf{b}^T_{\omega',\nu'}\right)^{1/2}\right]. \tag{201}$$

Here we can note that:

$$(\mathbf{b}_{\omega,\nu}\mathbf{b}^T_{\omega',\nu'})_{\eta k_2+\epsilon,\eta' k_2+\epsilon'} = \delta_{\eta,\omega}\delta_{\epsilon,\nu}\delta_{\eta',\omega'}\delta_{\epsilon',\nu'}. \tag{202}$$

We can then define two matrices:

$$(\mathbf{U} \in \mathbb{R}^{D_1D_2\times D_1D_2}) = \frac{1}{k_1k_2}\sum_{\omega,\omega'}^{k_1}\sum_{\nu,\nu'}^{k_2}(\mathbf{S}^T\mathbf{S})_{\omega k_2+\nu,\omega' k_2+\nu'}\tilde{\mathbf{A}}^T_{\omega,\nu}\tilde{\mathbf{A}}_{\omega,\nu}$$
$$(\mathbf{V} \in \mathbb{R}^{k_1k_2\times k_1k_2})_{\eta k_1+\nu,\eta' k_2+\nu'} = \frac{1}{D_1D_2}\mathrm{Tr}[\tilde{\mathbf{T}}\tilde{\mathbf{T}}\tilde{\mathbf{A}}^T_{\eta,\epsilon}\mathbf{A}_{\eta',\epsilon'}] \tag{203}$$

where $\tilde{\mathbf{T}} \otimes \mathbf{I}_n = \mathbf{T}$. We finally assume that $\boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}} \otimes \mathbf{I}_n$, where $\tilde{\boldsymbol{\Sigma}} \in \mathbb{R}^{D_1D_2\times D_1D_2}$. We can then find the following optimization:

$$\min_{\mathbf{S}\in\mathbb{R}^{k_1k_2\times k_1k_2}} \max_{\tilde{\mathbf{T}}\in\mathbb{R}^{D_1D_2\times D_1D_2}} \frac{1}{2k_1k_2}\boldsymbol{\theta}^T\boldsymbol{\Lambda}\boldsymbol{\theta}^* - \frac{1}{2D_1D_2}\mathrm{Tr}[\tilde{\mathbf{T}}^T\tilde{\mathbf{T}}] + \frac{1}{2k_1k_2}\mathrm{Tr}[\mathbf{S}^T\mathbf{S}\boldsymbol{\Lambda}]$$
$$+\frac{1}{D_1D_2}\mathrm{Tr}\left[\tilde{\mathbf{T}}\left(\tilde{\boldsymbol{\Sigma}} + \mathbf{U}\right)^{1/2}\right] - \frac{1}{k_1k_2}\mathrm{Tr}\left[\mathbf{S}\left(\boldsymbol{\Lambda}\boldsymbol{\theta}^*\boldsymbol{\theta}^{*T}\boldsymbol{\Lambda} + \frac{k_1k_2}{nD_1D_2}\mathbf{V}\right)^{1/2}\right]. \tag{204}$$

Similarly, for the case that contains a general regularization function, rather than the specific choice of the square regularizer, we can find the following form:

$$\min_{\mathbf{Q},\mathbf{U}\in\mathcal{S}_+^{k_1k_2}} \max_{\mathbf{P},\mathbf{V}\in\mathcal{S}_+^{k_1k_2}} \frac{1}{2}\mathrm{Tr}\left[\frac{1}{k_1k_2}\mathbf{U}\mathbf{P} - \mathbf{V}\mathbf{Q}\right] - \frac{1}{2nD_1D_2}\mathrm{Tr}\left[\mathbf{P}\mathbf{V}^{-1}\right] + \frac{\sigma_{\boldsymbol{\nu}}^2}{2D_1D_2}\mathrm{Tr}\left[(\mathbf{I}_{D_1D_2}+\mathbf{D})^{-1}\right]$$
$$+\frac{1}{k_1k_2}\mathbb{E}_{\mathbf{g}}\mathcal{M}_{\mathbf{V}\ R}\left(\boldsymbol{\theta}^* - \sqrt{\frac{k_1k_2}{nD_1D_2}}\sum_{\omega=1}^{k_1}\sum_{\nu=1}^{k_2}g_{\omega,\nu}\mathbf{V}^{-1}\mathbf{b}_{\omega,\nu}\right)$$
$$+\frac{1}{2D_1D_2}\mathrm{Tr}\left[\left(\sum_{\omega,\omega'=1}^{k_1}\sum_{\nu,\nu'=1}^{k_2}Q_{\omega k_2+\nu,\omega' k_2+\nu'}\tilde{\mathbf{A}}^T_{\omega,\nu}\tilde{\mathbf{A}}_{\omega',\nu'}\right)(\mathbf{I}_{D_1D_2}+\mathbf{D})^{-1}\right]. \tag{205}$$

Here $\mathbf{g} \in \mathbb{R}^{k_1k_2}$ is a zero mean Gaussian vector with covariance $\mathbb{E}[g_{\omega k_2+\nu}g_{\omega' k_2+\nu'}] = P_{\omega k_2+\nu,\omega' k_2+\nu'}$ for $\omega, \omega' \in [k_1], \nu, \nu' \in [k_2]$, and $\mathbf{D} \in \mathbb{R}^{D_1D_2\times D_1D_2}$ is defined by:

$$\mathbf{D} = \frac{1}{k_1k_2}\sum_{a,a'=1}^{k_1}\sum_{b,b'=1}^{k_2}U_{ak_2+b,a'k_2+b'}\tilde{\mathbf{A}}^T_{a,b}\tilde{\mathbf{A}}_{a',b'}. \tag{206}$$

G Numerical Simulation Details

For figures 1, 2, and 3, we compare the expected generalization error for both the primary and CGMT alternative optimizations.

For the primary optimization we make use of CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018) to directly solve to the optimization for Gaussian data $\mathbf{x}$ and labels $\mathbf{y}$ generated according to the rule specified in sections 5. Each primary optimization was repeated 100 times for different instantiations of the Gaussian data, and were averaged to approximate the expected value.

For the alternative optimizations, we do not make use of CVXPY, but instead solve optimizations by means of a fixed point iteration strategy, similarly to (Loureiro et al., 2021). As an example we will consider the specific case of the vector valued regression. Recalling equation (29), we can note that this optimization can be equivalently expressed as:

$$\min_{\mathbf{Q},\mathbf{U}\in\mathcal{S}_+^K}\max_{\mathbf{P},\mathbf{V}\in\mathcal{S}_+^K}\frac{1}{2mK}\mathrm{Tr}[\mathbf{\Lambda\Theta}^{*T}\mathbf{\Theta}^*]+\frac{1}{2K}\mathrm{Tr}[\mathbf{Q\Lambda}-\mathbf{P}]+\frac{1}{2K}\mathrm{Tr}[\mathbf{U}^{-1}\mathbf{P}-\mathbf{V}^{-1}\mathbf{Q}]+\frac{1}{2K}\mathrm{Tr}[\mathbf{U}(\mathbf{\Sigma}+\mathbf{Q})]$$
$$-\frac{1}{2K}\mathrm{Tr}\left[\mathbf{V}\left(\frac{m}{n}\mathbf{P}+\frac{1}{m}\mathbf{\Lambda\Theta}^{*T}\mathbf{\Theta}^*\mathbf{\Lambda}\right)\right]. \quad (207)$$

We can note this equivalence by solving over $\mathbf{V}$ and $\mathbf{U}$, and then defining $\mathbf{S},\mathbf{T}\in\mathcal{S}_+^T$ such that $\mathbf{Q}=\mathbf{S}^T\mathbf{S}$ and $\mathbf{P}=\mathbf{T}^T\mathbf{T}$. By taking the derivatives with respect to the four parameters $\mathbf{Q},\mathbf{U},\mathbf{P},\mathbf{V}$ we can find the following set of equations:

$$\mathbf{U} = \left[\tfrac{m}{n}\mathbf{V}+\mathbf{I}_K\right]^{-1}, \quad (208)$$
$$\mathbf{V} = \left[\mathbf{U}+\mathbf{\Lambda}\right]^{-1}, \quad (209)$$
$$\mathbf{Q} = \mathbf{V}\left[\tfrac{m}{n}\mathbf{P}+\mathbf{\Lambda\Theta}^{*T}\mathbf{\Theta\Lambda}\right]\mathbf{V}, \quad (210)$$
$$\mathbf{P} = \mathbf{U}\left[\mathbf{\Sigma}+\mathbf{Q}\right]\mathbf{U}. \quad (211)$$

By starting with an initial guess of $\mathbf{U}=\mathbf{V}=\mathbf{P}=\mathbf{Q}=\mathbf{I}_K$ and repeatedly iterating this set of equations the values of $\mathbf{U},\mathbf{V},\mathbf{P},\mathbf{Q}$ converge rapidly to their unique solution. Furthermore as all matrices involved in the iteration are PSD, this invariant is maintained at all steps of the iterations, in contrast to gradient methods which require projection onto the PSD matrix cone at each step of the iteration. Similar fixed point equations can be readily found for the complex, and convolutional cases discussed in section 5.