In-Context Benign Overfitting: A Feature-Selection Model in Linear Regression ICL

Puneesh Deora
University of British Columbia
puneeshdeora@ece.ubc.ca

Bhavya Vasudeva
University of Southern California
bvasudev@usc.edu

Christos Thrampoulidis
University of British Columbia
cthrampo@ece.ubc.ca

Abstract

In in-context learning (ICL), a frozen pre-trained model solves tasks by conditioning on a prompt of a few input–output examples, without gradient updates. If the task was present in pretraining but the particular prompt sequence was unseen, the resulting in-distribution generalization is retrieval-based ICL. Learning-based ICL instead reflects out-of-distribution generalization, where the model succeeds on prompts generated by a novel task. Empirically, both forms improve with scale. By analogy to benign overfitting in supervised learning, we call this in-context benign overfitting: larger models more faithfully memorize the pretraining tasks (improving retrieval ICL) while also generalizing better to novel tasks (improving learning ICL). We prove that this phenomenon already arises in a minimal in-context linear-regression feature-selection model. In contrast, standard in-context linear-regression models exhibit a retrieval–learning tradeoff, where the emergence of learning-based ICL coincides with degraded retrieval-based performance.

1 Introduction

In-context learning (ICL) has established itself as a cornerstone capability of large language models, where the model solves tasks at inference time by conditioning on a prompt of input–output demonstrations, without any updates to model weights [Brown et al., 2020]. While the prompt contains only a handful of examples, the model also draws on information acquired during pretraining. As a result, there are two qualitatively different settings in which a model can succeed: it can perform well because the underlying task was present during pretraining, or it can perform well on prompts generated by a task that is novel relative to pretraining. Pan et al. [2023] introduced a useful distinction between these two modes of ICL, which they call *retrieval* and *learning*. Concretely, let

$$\mathcal{D}_{\text{task}}^{\text{train}} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_M\}$$

denote the set of tasks present during pretraining. For a task $\boldsymbol{w}$, a prompt consists of T demonstrations and a query,

$$X_{\boldsymbol{w}} = [(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_T, y_T), \boldsymbol{x}_q],$$

where labels $y_t = f_{\boldsymbol{w}}(\boldsymbol{x}_t)$ are generated through a fixed mapping f that is shared across demonstrations inside the prompt. This mapping is parameterized by the task $\boldsymbol{w}$ and applies to all $t \in [T]$ as well as the query $\boldsymbol{x}_q$. The two modes are then defined as follows:

- **Retrieval mode (ID generalization).** We sample $\boldsymbol{w} \sim \mathcal{D}_{\text{task}}^{\text{train}}$, so while the prompt $X_{\boldsymbol{w}}$ is unseen during pretraining, the underlying task is drawn from the pretraining task set. We evaluate whether the model correctly predicts the query label y_q. Success in this mode is consistent with the

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

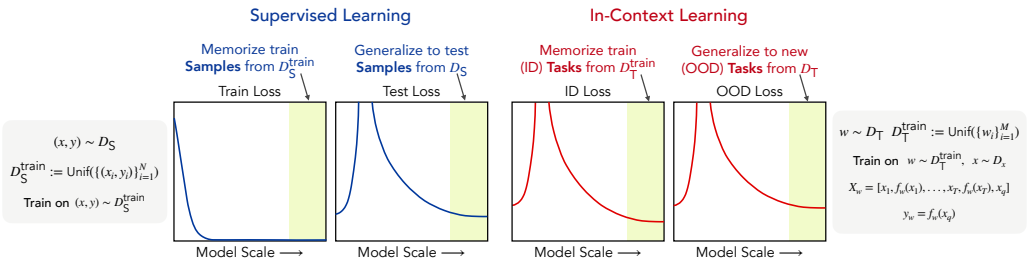


Figure 1: **Benign Overfitting in Supervised vs. In-Context Learning (Cartoon Illustration).** Left illustrates benign overfitting commonly observed in standard supervised learning setups, where large models far beyond the interpolation threshold (green band) memorize the train *samples* while simultaneously generalizing to new ones. We introduce and formalize an in-context analogue of this phenomenon (right): large models (in green) memorize *tasks* seen during (pre)training while simultaneously generalizing to novel out-of-distribution tasks.

model having *memorized* the pretraining tasks and *retrieving* the correct task mechanism from its parameters. This mode, therefore, probes task-level in-distribution (ID) generalization.

- **Learning mode (OOD generalization).** We sample $\boldsymbol{w} \sim \mathcal{D}_{\text{task}}$ from a distribution that differs from $\mathcal{D}_{\text{task}}^{\text{train}}$. We again evaluate prediction on y_q. Since the task is novel relative to pretraining, the model cannot rely on task memorization alone and must *learn* from demonstrations. This mode probes task-level out-of-distribution (OOD) generalization.

Hereon, we will refer to these as ID and OOD generalization, respectively.

Understanding these two modes has motivated a surge of work in synthetic, fully controlled settings. These studies train transformers from scratch on sequences generated by a finite task set, for example in-context linear regression [Garg et al., 2022, Raventos et al., 2023, Carroll et al., 2025] or Markov-chains [Park et al., 2025], and then evaluate ID and OOD generalization as discussed above. This line of work has uncovered several qualitative phenomena, including task-diversity thresholds [Raventos et al., 2023, Park et al., 2025] for the emergence of *good* OOD generalization and transient dynamics [Singh et al., 2023, Carroll et al., 2025] in how the two modes evolve over the course of training. We defer a detailed discussion and pointers to the relevant papers to Appendix A.

One question, however, remains conceptually unsettled: *how should ID and OOD generalization evolve with model scale?* In the era of scaling laws, the prevailing intuition is that larger models should excel in both modes. Indeed, this has been observed experimentally: Wei et al. [2023] (Figs. 2 and 3) show that in large language models, both modes tend to improve with scale: larger models more accurately solve tasks encountered during pretraining and also generalize better to novel tasks at inference time.

Contributions. Drawing an analogy to the literature on benign overfitting in supervised learning [Bartlett et al., 2020, Hastie et al., 2019, Belkin et al., 2020], we frame the scaling behavior of ID and OOD generalization in ICL as what we call *in-context benign overfitting*.

In-Context Benign Overfitting

Definition 1. *In-context benign overfitting is the property whereby large (overparameterized) models memorize ID ICL tasks, achieving low risk on prompts from $\mathcal{D}_{task}^{train}$, while simultaneously generalizing to novel OOD ICL tasks (from $\mathcal{D}_{task}$).*

This lifts the classical notion of benign overfitting from the sample level to the task level: whereas standard benign overfitting concerns memorization of noisy training *samples*, in-context benign overfitting concerns memorization of pretraining *tasks*. See Fig. 1 for an illustration.

We ground our framework in the minimal setup of in-context linear regression [Garg et al., 2022], employing a simplified linear attention model consistent with recent theoretical works [Zhang et al., 2023, Lu et al., 2025, Wu et al., 2023, Zhang et al., 2025]. To explicitly model the effect of scale, we incorporate a feature subsampling mechanism inspired by the double descent literature [Hastie et al., 2019, Belkin et al., 2020]. Using data generated from this model, we study the two canonical

predictors extensively studied in classical statistical learning: the minimum-norm interpolator (in the overparameterized regime) and the least-squares solution (in the underparameterized regime). We analyze these predictors in the proportional high-dimensional asymptotic limit, where the number of (pre)training tasks, number of (pre)training prompts, number of in-context demonstrations, number of learnable parameters, and the ambient feature dimension grow strictly in proportion [Lu et al., 2025]. In this regime, we derive sharp analytical characterizations of the ID and OOD risks, demonstrating that our theoretical predictions accurately track the empirical risk curves and faithfully capture the phenomenon of in-context benign overfitting. Beyond capturing in-context benign overfitting, our framework is consistent with prior literature [Raventos et al., 2023, Wu et al., 2023], faithfully reproducing the known effects of task diversity on ID and OOD generalization.

2 Problem Setup

Data Model. We study in-context linear regression first introduced by Garg et al. [2022]. In this setup, input sequences to the model are interleaved input-output pairs $(\boldsymbol{x}, y)$ of the form $X = [\boldsymbol{x}_1, y_1, \boldsymbol{x}_2, y_2, \ldots, \boldsymbol{x}_T, y_T, \boldsymbol{x}_{T+1}]$, with feature vectors $\boldsymbol{x}_t \in \mathbb{R}^d$ and labels $y_t \in \mathbb{R}$. The feature vectors $\boldsymbol{x}_t$ are sampled i.i.d from a Gaussian distribution, *i.e.*, $\boldsymbol{x}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$. Each sequence has an associated *task vector* $\boldsymbol{w} \sim \mathcal{D}_{\text{task}} = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ that generates labels

$$y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon_t, \quad \forall t \in [T+1],$$

with noise $\epsilon_t \sim \mathcal{N}(0, \sigma_n^2)$. We refer to $\boldsymbol{x}_q := \boldsymbol{x}_{T+1}$ as the query vector, $y_q := y_{T+1}$ as the query label, and the number of demonstrations T as the context-length. Given the T demonstrations $(\boldsymbol{x}_t, y_t)_{t=1}^T$, the goal is predicting y_q.

Training Data. To construct the training set, we first sample a finite pool of M task vectors $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M \sim \mathcal{D}_{\text{task}}$ independently, and define the train task distribution as uniform over this pool:

$$\mathcal{D}_{\text{task}}^{\text{train}} = \text{unif}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}.$$

We then generate N training sequences. For each $i \in [N]$, we sample $\boldsymbol{w}_i \sim \mathcal{D}_{\text{task}}^{\text{train}}$ and construct

$$X_i = [\boldsymbol{x}_1^i, y_1^i, \boldsymbol{x}_2^i, y_2^i, \ldots, \boldsymbol{x}_T^i, y_T^i, \boldsymbol{x}_q^i],$$

where $\boldsymbol{x}_t^i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $y_t^i = \boldsymbol{w}_i^\top \boldsymbol{x}_t^i + \epsilon_t^i$ with $\epsilon_t^i \sim \mathcal{N}(0, \sigma_n^2)$ for all $t \in [T+1]$.

Learning Model. To solve the ICL linear regression problem, we employ a simplified model inspired by linear attention, augmented with a feature-selection mechanism designed to study the phenomenon of in-context benign overfitting. In its vanilla form (without feature selection), this model has been used extensively in prior work as a tractable framework for theoretically analyzing ICL [Zhang et al., 2025, Lu et al., 2025, Wu et al., 2023, Zhang et al., 2024, Frei and Vardi, 2024]; see Zhang et al. [2025] for details on the reduction from linear attention.

Since we want to study how model scale affects ID (retrieval) and OOD generalization (learning), we need a knob for model capacity. Our key modeling contribution is a feature-selection formulation inspired by the double descent literature in supervised learning [Hastie et al., 2019, Belkin et al., 2020], which exactly captures this. This formulation traces its roots to classical statistics [Breiman and Freedman, 1983]. In analogy to that literature, we introduce here feature selection as a minimal mechanism sufficient to demonstrate in-context benign overfitting. Specifically, we vary model capacity by restricting access to a subset $\mathcal{S} \subseteq \{1, \ldots, d\}$ of the original features. For each feature vector $\boldsymbol{x}_t$, the learner observes only the subvector $\boldsymbol{x}_t[\mathcal{S}]$, where for any vector $\boldsymbol{v}$ we use $\boldsymbol{v}[\mathcal{S}]$ to denote its restriction to the indices in $\mathcal{S}$. We denote the number of selected features by $p = |\mathcal{S}|$, and select $\mathcal{S}$ by sampling p indices uniformly without replacement from $[d]$.

Given $\mathcal{S}$, the model uses the T demonstrations $\{(\boldsymbol{x}_t[\mathcal{S}], y_t)\}_{t\in[T]}$ and the query subvector $\boldsymbol{x}_q[\mathcal{S}]$ to construct the feature embedding

$$\mathbf{h}_{\mathcal{S}_p}(X) = \texttt{vec}\big(\boldsymbol{x}_q[\mathcal{S}]\, \hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}]^\top\big) \in \mathbb{R}^{p^2}, \tag{1}$$

where $\hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}] = \frac{1}{T}\sum_{t=1}^T y_t \boldsymbol{x}_t[\mathcal{S}]$ is the *averaging estimator* restricted to $\mathcal{S}$. The model prediction is then linear in these features:

$$f_{\boldsymbol{\theta},\mathcal{S}}(X) := \boldsymbol{\theta}^\top \mathbf{h}_{\mathcal{S}_p}(X) = \boldsymbol{x}_q[\mathcal{S}]^\top \boldsymbol{\Theta}\, \hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}],$$

where the trainable parameter $\boldsymbol{\theta} \in \mathbb{R}^{p^2}$ is the vectorization of $\boldsymbol{\Theta} \in \mathbb{R}^{p\times p}$, i.e., $\boldsymbol{\theta} = \texttt{vec}(\boldsymbol{\Theta})$. The model has p^2 trainable parameters, for simplicity, we refer to p as the *model size* throughout. The vanilla formulation of prior work [Zhang et al., 2025, Lu et al., 2025, Zhang et al., 2024] corresponds to $p = d$. To see why this is a sensible predictor, observe that $\hat{\boldsymbol{w}}_{\text{avg}}$ provides an estimate of the task vector $\boldsymbol{w}$. Indeed, writing $\hat{\boldsymbol{w}}_{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} y_t \boldsymbol{x}_t = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon_t)\boldsymbol{x}_t$, and noting that $\mathbb{E}[\epsilon_t \boldsymbol{x}_t] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{x}_t \boldsymbol{x}_t^\top] = \boldsymbol{\Sigma}_{\boldsymbol{x}}$, we have $\mathbb{E}[\hat{\boldsymbol{w}}_{\text{avg}}] = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}$. Thus, when $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \sigma_x^2 \mathbb{I}_d$, the averaging estimator is an unbiased estimator of $\sigma_x^2 \boldsymbol{w}$.

Training Algorithm. Given the training sequences $\{X_i\}_{i=1}^N$ and a model of size p acting on an associated feature subset $\mathcal{S}$ with $|\mathcal{S}| = p$, let $\boldsymbol{H}_{\mathcal{S}_p} = [\mathbf{h}_{\mathcal{S}_p}(X_1), \ldots, \mathbf{h}_{\mathcal{S}_p}(X_N)]^\top \in \mathbb{R}^{N\times p^2}$ denote the feature matrix and $\boldsymbol{y} = [y_q^1, \ldots, y_q^N]^\top \in \mathbb{R}^N$ the vector of query labels. The parameter $\boldsymbol{\theta}$ is obtained by minimizing the empirical risk under squared loss:

$$\mathcal{L}_{\text{emp}}(\boldsymbol{\theta}) := \frac{1}{N}\sum_{i=1}^{N}\big(\boldsymbol{\theta}^\top \mathbf{h}_{\mathcal{S}_p}(X_i) - y_q^i\big)^2 = \frac{1}{N}\|\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{\theta} - \boldsymbol{y}\|^2.$$

The limiting solution $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ (when training with GD) depends on the problem geometry. In the *underparameterized regime*, if $\boldsymbol{H}_{\mathcal{S}_p}$ has full column rank, $\mathcal{L}_{\text{emp}}$ is strongly convex and GD converges to the unique least-squares solution $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p} := \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{\theta}\|^2$. In the *overparameterized regime*, if $\boldsymbol{H}_{\mathcal{S}_p}$ has full row rank, there exist infinitely many interpolators, and GD initialized at zero exhibits an implicit bias towards the minimum-norm solution [Hastie et al., 2019, Bartlett et al., 2020], converging to $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p} = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|$ s.t. $\mathbf{H}_{\mathcal{S}_p}\boldsymbol{\theta} = \mathbf{y}$. This characterization allows us to study the generalization properties of these limiting predictors as the model size p varies.

Generalization Metrics. We evaluate the trained predictor $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ on two types of generalization: (i) **In-distribution (ID)**: performance on training tasks, i.e., $\boldsymbol{w} \sim \mathcal{D}_{\text{task}}^{\text{train}}$; (ii) **Out-of-distribution (OOD)**: performance on novel tasks, i.e., $\boldsymbol{w} \sim \mathcal{D}_{\text{task}}$.

To define these metrics, we first introduce the *task-specific mean-square-error loss*, which measures the prediction error of $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ on sequences generated from a fixed task vector $\boldsymbol{w}$: $\mathcal{L}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}; \boldsymbol{w}) := \mathbb{E}_{X,\epsilon}\big[(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbf{h}_{\mathcal{S}_p}(X) - y_q)^2\big]$, where the expectation is over the feature vectors and noise in the sequence X, i.e., $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \boldsymbol{x}_q \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$ and $\epsilon_1, \ldots, \epsilon_q \sim \mathcal{N}(0, \sigma_n^2)$. We use the shorthand $\mathbb{E}_{X,\epsilon}$ for this expectation throughout.

The ID and OOD risks are then defined by averaging over the respective task distributions:

$$\mathcal{L}_{\text{ID}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) := \mathbb{E}_{\mathcal{S}_p}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{D}_{\text{task}}^{\text{train}}}\mathcal{L}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}; \boldsymbol{w}), \quad \mathcal{L}_{\text{OOD}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) := \mathbb{E}_{\mathcal{S}_p}\mathbb{E}_{\boldsymbol{w}\sim\mathcal{D}_{\text{task}}}\mathcal{L}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}; \boldsymbol{w}). \tag{2}$$

In both cases, the outer expectation $\mathbb{E}_{\mathcal{S}_p}$ averages over the random feature subset $\mathcal{S}$, ensuring that we evaluate performance independently of any particular realization of it. We note that this notion of OOD generalization differs from traditional distribution shift: as $M \to \infty$, the training task distribution $\mathcal{D}_{\text{task}}^{\text{train}}$ converges to $\mathcal{D}_{\text{task}}$, so the gap is finite-$M$ in nature. This terminology follows prior work [Park et al., 2025], where OOD refers to generalization to tasks not seen during pretraining.

3 Sharp Asymptotics of ID and OOD risks

3.1 Asymptotic setting

We characterize the ID and OOD risks in the joint high-dimensional limit where the ambient dimension d, context length T, model size p, number of training tasks M, and number of training sequences N all grow to infinity as follows: $N, T, M, d, p \to \infty$ with

$$\frac{N}{d^2} \to \nu, \quad \frac{M}{d} \to \mu, \quad \frac{T}{d} \to \tau, \quad \frac{p}{d} \to \rho. \tag{3}$$

To build intuition for these scalings, the ratios ν, μ, τ, ρ arise naturally from the problem structure. Specifically, as $d \to \infty$, $\rho = p/d$ captures model scale, $\mu = M/d$ normalizes task diversity since the empirical task covariance concentrates around $\mathbb{I}_d$ when $M = \Theta(d)$. The ratio $\tau = T/d$ normalizes context length since, as discussed in Section 2, $\hat{\boldsymbol{w}}_{\text{avg}}$ is an unbiased estimator of $\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{w}$, and with $T = \Theta(d)$ demonstrations it concentrates around this value, giving accurate predictions for the query

$\boldsymbol{x}_q$. Finally, given the feature matrix $\boldsymbol{H}_{\mathcal{S}_p} \in \mathbb{R}^{N \times p^2}$, the interpolation threshold is $p^2 = N$. Since $p \propto d$, we naturally normalize the number of training sequences N by d^2, giving $\nu = N/d^2$ as the scaling, and $\rho^2 = \nu$ as the interpolation threshold. Here, $\nu > 0, \mu > 0, \tau > 0$ and $\rho \in (0, 1]$ are constants that as we will see determine the ID/OOD risk asymptotics.

Throughout this section, we assume isotropic features and task vectors, i.e., $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \sigma_x^2 \mathbb{I}_d$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \sigma_w^2 \mathbb{I}_d$. As we will show this setting already suffices to establish the in-context benign overfitting phenomenon. Similar to the setup of Lu et al. [2025], we fix $\sigma_x = 1/\sqrt{d}, \sigma_w = 1$. We defer extensions to general covariances to future work.

Our goal is to compute the in-probability asymptotic limits $\operatorname{plim}_{d\to\infty} \mathcal{L}_{\text{ID}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p})$ and $\operatorname{plim}_{d\to\infty} \mathcal{L}_{\text{OOD}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p})$ under the scaling in Eq. (3). Here, $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ is given by LS or min-norm interpolators defined in Section 2.

3.2 Linear Gaussian Equivalent Model

Our asymptotic analysis leverages powerful machinery from high-dimensional statistics that provides sharp characterizations for solutions of *Linear Gaussian Problems* (LGPs).

Formally, an $\text{LGP}(\boldsymbol{\theta}_*, \boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_\epsilon)$ is a linear regression problem with isotropic Gaussian features $\boldsymbol{g}_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{\tilde{p}})$, labels $y_i = \boldsymbol{g}_i^\top \boldsymbol{\Sigma}_g^{1/2} \boldsymbol{\theta}_* + \epsilon_i$ with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$. Analogously to Section 2, given N samples $\{(\boldsymbol{g}_i, y_i)\}_{i\in[N]}$, for this LGP, we define the estimator $\hat{\boldsymbol{\theta}}$ as the LS solution when $\tilde{p} < N$ and the min-norm interpolator when $\tilde{p} \geq N$; see Appendix B for the full definition.

LGPs are central to our analysis for two reasons: (i) Well-established tools yield sharp asymptotic characterizations for the generalization error of squared-loss minimizers in LGPs; (ii) A *Gaussian equivalence principle* allows us to transfer these LGP results to our original (non-Gaussian) problem setup in Section 2. We will empirically verify the accuracy of these predictions.

The LGP solves LS or minimum-norm optimization as in LS and minimum-norm solutions of our original setup. However, there are two key advantages in the LGP setting. First, the transition threshold between regimes is precisely $\tilde{p} = N$: as $\tilde{p}$ and N grow proportionally, the Gaussian feature matrix $\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_N]^\top \in \mathbb{R}^{N \times \tilde{p}}$ is almost surely full column rank if and only if $\tilde{p} < N$. Second, and most importantly, the optimization problems (LS or min-norm) for LGP involve only an isotropic Gaussian matrix and independent Gaussian noise, which admits sharp asymptotic characterizations. This is in contrast to the the original problems in Section 2 involve non-Gaussian features $\mathbf{h}_{\mathcal{S}_p}(X_i)$ (Eq. (1)) and correlated labels $y_q^i = \boldsymbol{w}_i^\top \boldsymbol{x}_q + \epsilon$. Next, we explicitly relate LGPs to our original setup of interest detailed in Section 2. Since the feature $\mathbf{h}_{\mathcal{S}_p}(X_i) \in \mathbb{R}^{p^2}$, the regressor dimension is $\tilde{p} = p^2$.

Definition 2 (Linear Gaussian Equivalent Problem)**.** *Consider N features $\mathbf{h}_{\mathcal{S}_p}(X_i) \in \mathbb{R}^{p^2}$ as in Eq. (1) with labels $y_q^i = \boldsymbol{w}_i^\top \boldsymbol{x}_q + \epsilon$ for $\boldsymbol{w}_i \sim \mathcal{D}_{\text{task}}^{\text{train}}$. The equivalent $\text{LGP}(\boldsymbol{\theta}_*, \boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_\epsilon)$ is given by setting $\boldsymbol{\Sigma}_g \leftarrow \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$, $\boldsymbol{\theta}_* \leftarrow \frac{1}{N} \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_\epsilon \leftarrow \text{diag}(\sigma_{n_1}, \ldots, \sigma_{n_N})$ with*

$$\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} := \frac{1}{N} \mathbb{E}_{X,\epsilon}\left[\boldsymbol{H}_{\mathcal{S}_p}^\top \boldsymbol{H}_{\mathcal{S}_p}\right] \in \mathbb{R}^{p^2 \times p^2}, \tag{4a}$$

$$\boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}} = \mathbb{E}_{X,\epsilon}[\boldsymbol{H}_{\mathcal{S}_p}^\top \boldsymbol{y}] \tag{4b}$$

$$\sigma_{n_i}^2 := \sigma_x^2 \|\boldsymbol{w}_i\|^2 + \sigma_n^2 - \boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} \boldsymbol{\theta}_\star \; i \in [N]. \tag{4c}$$

See App. B for explicit formulas for the expectations above.

Equivalence is in the strong sense that the asymptotic *empirical distribution* of the LGP estimator $\hat{\boldsymbol{\theta}}$ matches that of the original estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}}$ [Goldt et al., 2020, Hu and Lu, 2020]. Thus, it suffices to characterize the former, since knowing the latter suffices to compute ID and OOD limiting risks.

The LGP parameters are chosen so that the first and second-order statistics of the LGP samples $\{(\boldsymbol{g}_i, y_i)\}_{i\in[N]}$ match the first and second-order statistics (with respect to X, ϵ) of the original data $\{(\mathbf{h}_{\mathcal{S}}(X_i), y_q^i)\}_{i\in[N]}$. The principle that Gaussian systems matching a non-Gaussian system to first and second order are predictive of properties of the latter is broadly known as *universality*, with roots in classical random matrix theory [Tao, 2023]. The specific form relevant here, where the "system" refers to a random optimization problem and "properties" to quantities such as its optimal cost or empirical distribution of its solution, has been developed within the high-dimensional statistics

literature, from early applications in compressed sensing [Montanari and Nguyen, 2017, Oymak and Tropp, 2018] to more recent machine learning settings [Hastie et al., 2019, Hu and Lu, 2022]. The terminology *Gaussian equivalence principle* (GEP) was introduced by Mei and Montanari [2019], Goldt et al. [2020], and since then numerous works have identified and confirmed this equivalence in various contexts; we refer the reader to App. A for a detailed discussion. Here, we identify another instance where the GEP applies; Fig. 2 empirically demonstrates this. Formalizing this rigorously is left for future work, as our focus here is on deriving predictions and establishing in-context benign overfitting. A formal proof could likely build on recent universality equivalence proofs [Hu and Lu, 2022, Montanari et al., 2023]. We state our main results below.

3.3 Asymptotics of the Gaussian Equivalent Problem

Theorem 1 (Asymptotic Risk of Minimum-Norm LGP)**.** *Consider the minimum norm estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ for the linear Gaussian equivalent problem in Definition 2 under the overparameterized regime ($\nu < \rho^2$). Under the asymptotic scaling defined in Eq. (3), as $d \to \infty$, the ID and OOD risks converge in probability to the following deterministic limits:*

$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{OOD}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = \left(1 - \tfrac{\nu}{\rho^2}\right) \rho(1+ac)(1-c(1+a+ac)) + (1+a+ac)(\sigma_n^2 + \rho c + 1 - \rho),$$
$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{ID}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = \frac{\nu}{a^2}\bar{\beta}_*^2,$$

where $c := \frac{\sigma_n^2+1}{\tau}$ and the scalars a and $\bar{\beta}_*$ are the unique positive solutions to Eq. (8).

Theorem 2 (Asymptotic Risk of LS LGP)**.** *Consider the least-squares estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ for the linear Gaussian equivalent problem in the underparameterized regime ($\nu > \rho^2$). Under the asymptotic scaling in Eq. (3), as $d \to \infty$, the ID and OOD risks converge in probability as:*

$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{OOD}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = m_c\big(\kappa_\infty^2(1+c) - 2c^2\rho\big) + (1+c)\rho c^2 m_c' + \sigma_n^2 + 1 - \rho + \rho c,$$
$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{ID}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = \frac{\nu}{\rho^2}\kappa_\infty^2,$$

where the scalar $\kappa_\infty := \sqrt{\frac{\sigma_n^2+1-\rho+\rho c-\rho c^2 m_c}{\frac{\nu}{\rho^2}-1}}$, $c := \frac{\sigma_n^2+1}{\tau}$, and $m_c := m_\gamma(z_c)$ is the Stieltjes transform of the Marchenko-Pastur law with parameter $\gamma = \rho/\mu$ evaluated at $z_c = -c$. In addition, m_c' denotes the derivative of $m_\gamma(\cdot)$ at z_c.

Theorems 1 and 2 together provide sharp closed-form characterizations of the ID and OOD risks across both regimes, in terms of the key problem ratios ν, τ, μ, ρ. While the expressions may appear complex, they are explicit and computable given the problem parameters. Notably, from the closed-form expression of κ_∞, it is immediately clear that $\kappa_\infty \to \infty$ as $\nu/\rho^2 \to 1^+$, confirming that both risks diverge at the interpolation threshold $\rho^2 = \nu$. The scalars a and $\bar{\beta}_*$, which encode the effect of model scale ρ^2/ν on the risks implicitly, arise as solutions to a low-dimensional fixed-point system, and the Stieltjes transform of the Marchenko-Pastur law $m_\gamma(\cdot)$, which captures the spectral structure of the task covariance, arises naturally from the problem geometry. We discuss how they emerge in Sec. 5. We next verify empirically that the theoretical predictions of Thm. 1 and 2 closely match observed risk curves, and demonstrate the phenomenon of in-context benign overfitting.

4 Experimental Results

We consider the same setup as described in Section 2, unless stated otherwise. See Appendix H for details on the experimental settings.

Model Scale and Benign Overfitting. Fig. 2 compares our derived asymptotic risk curves (Theorems 1 and 2) against empirical loss values. We plot both ID and OOD risk as a function of the relative model scale ρ^2/ν (specifically, we fix N and vary p).

We observe that the theoretical predictions closely match the empirical observations. The vertical asymptote at $\rho^2/\nu = 1$ marks the phase transition between the underparameterized ($\rho^2/\nu < 1$) and overparameterized ($\rho^2/\nu > 1$) regimes (as discussed in Section 3.3). As model scale increases, both ID and OOD loss exhibit double descent. Importantly, increasing the model scale causes the ID loss to converge to a significantly lower value compared to the underparameterized regime, while the OOD loss converges to a value comparable to the underparameterized minimum. This demonstrates that the

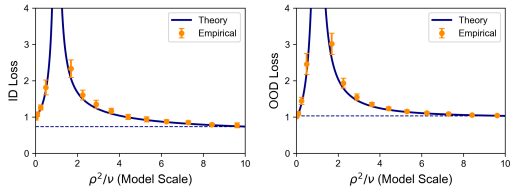


Figure 2: **In-Context Benign Overfitting**. Comparison of our theoretical risk curves (solid lines) vs. empirical observations (points) for ID Loss (left) and OOD Loss (right) as a function of model scale ρ^2/ν (with fixed ν). The dashed lines denote the asymptotic value in the overparameterized regime. The vertical asymptote at $\rho^2/\nu = 1$ marks the interpolation threshold. In the overparameterized regime ($\rho^2/\nu > 1$), the model exhibits in-context benign overfitting: it effectively memorizes pre-training tasks (low ID risk) while simultaneously yielding a small risk on OOD tasks.

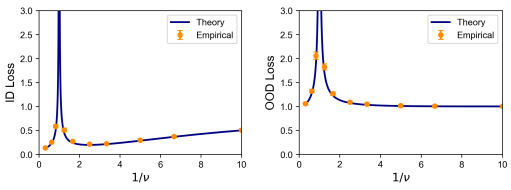


Figure 3: **Absence of In-Context Benign Overfitting.** Theoretical (lines) vs. empirical (points) ID (left) and OOD (right) risk as a function of overparameterization $1/\nu$ with fixed $\rho = 1$ (recovering the setting of Lu et al. [2025]). In contrast to Fig. 2, this setting exhibits a retrieval-learning tradeoff: as the model becomes more overparameterized, OOD generalization improves, but ID performance degrades (higher loss). This underscores that varying ambient dimension d is distinct from scaling model capacity p against a fixed data distribution.

model benignly overfits the training tasks—memorizing ID tasks while simultaneously generalizing to OOD tasks, capturing the essence of in-context benign overfitting.

Comparison with Lu et al. [2025]. Next, in Fig. 3, we isolate the setting studied by Lu et al. [2025] by fixing $\rho = 1$ (i.e., using all features) and vary the overparameterization ratio via $1/\nu$. We visualize the theoretical and empirical ID (left) and OOD (right) loss values. Crucially, in this setting, we do not observe in-context benign overfitting. While the OOD loss decreases as expected, the ID loss monotonic increases with overparameterization. This divergence stems from the nature of the scaling: varying $1/\nu$ requires altering the ambient dimension d relative to N, which fundamentally changes the underlying data distribution. In contrast, our feature-selection model (varying $\rho < 1$) allows us to scale the model capacity p while keeping the underlying data distribution fixed—a proxy that we argue more faithfully captures the effect of scaling model size. Consistent with our perspective, this parallels observations in supervised learning: standard linear regression without feature selection does not exhibit benign overfitting [Hastie et al., 2019].

Impact of Task Diversity and Scale. Next, we investigate the role of varying the number of training tasks M, which we call task diversity μ and its interplay with model scale ρ. Figure 4 visualizes our theoretically derived ID (left) and OOD (right) risk landscapes as a joint function of model scale (ρ) and task diversity (μ). Our analysis recovers and clarifies the task-diversity thresholds identified in prior work [Raventos et al., 2023, Park et al., 2025]:

Fixed Task Diversity (μ): Consistent with Fig. 2, for any fixed number of pre-training tasks, increasing model scale (ρ) induces double descent in both ID and OOD risks. Consequently, sufficiently large models ($\rho \approx 1$) achieve lower ID loss than their underparameterized counterparts, and similar OOD loss confirming that in-context benign overfitting persists across different diversity regimes.

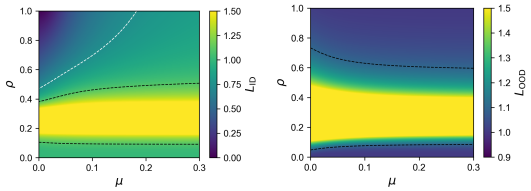
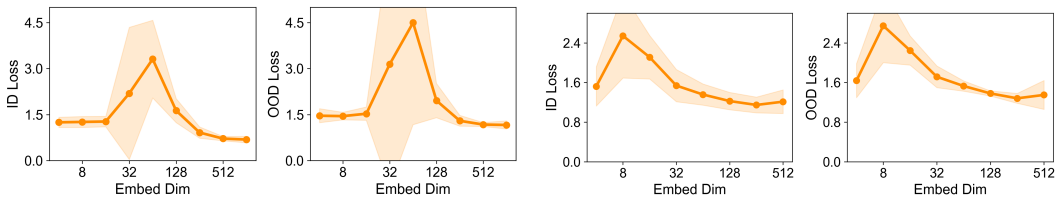


Figure 4: **Task Diversity vs. Model Scale**. Theoretical heatmaps for ID Loss (left) and OOD Loss (right) as a function of task diversity μ and model scale ρ. Lighter colors indicate higher loss. Vertical cuts (fixed μ) demonstrate in-context benign overfitting: larger models ($\rho \approx 1$) minimize ID loss without sacrificing OOD performance. Horizontal cuts (fixed ρ) reveal the task diversity threshold: increasing diversity forces a transition from memorization (low ID, high OOD) loss to learning-based ICL (high ID, low OOD risk).

Fixed scale (ρ): In contrast, for a fixed overparameterized model ($\rho > \sqrt{\nu}$), increasing task diversity (μ) presents a trade-off. Higher diversity improves OOD generalization but degrades ID performance, as memorizing more tasks becomes increasingly difficult for a fixed-capacity model. The critical



(a) Linear Attention only (b) Full transformer (softmax)

Figure 5: **Transformer Experiments.** ID (left) and OOD (right) loss as a function of embedding dimension for a 1 layer, 1 head transformer with (a) linear attention only; and (b) full transformer with softmax attention, trained on in-context linear regression. Both exhibit in-context benign overfitting; see App. H for train loss results illustrating interpolation thresholds.

point where ID performance begins to degrade while OOD performance improves corresponds to the task diversity threshold established in previous literature [Raventos et al., 2023, Park et al., 2025].

Transformer Experiments. To validate that in-context benign overfitting extends beyond our theoretical model, we train 1 layer transformer models on in-context linear regression. Figure 5 shows ID and OOD loss as a function of embedding dimension for both linear attention only and full transformer models with softmax attention. In both cases, we observe in-context benign overfitting, consistent with our theoretical predictions, with heavily overparameterized models simultaneously achieving the low ID and OOD risks. We refer to App. H for full experimental details, including ablations such as adding different model components and deeper models.

5 Proof Sketch

Below we provide the proof sketch to derive asymptotic characterizations for the Gaussian equivalent problem of Def. 2 for the overparameterized regime $\rho^2 > \nu$ (i.e., $p^2 > N$), and defer the underparameterized regime to App E. We begin by analyzing the following minimum-norm linear Gaussian equivalent optimization:

$$\min_{\boldsymbol{a}} \ \frac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_*\|^2 \quad \text{s.t.} \quad \boldsymbol{G}\boldsymbol{a} = \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}', \tag{Primal}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$ and $\boldsymbol{\theta}_*$ are as in Definition 2; $\boldsymbol{D}_{\boldsymbol{w}} \in \mathbb{R}^{N\times N}$ is a diagonal matrix with entries $(\boldsymbol{D}_{\boldsymbol{w}})_{ii} = \sigma_{n_i}$ (see Definition 2 for σ_{n_i}); $\boldsymbol{G} \in \mathbb{R}^{N\times p^2}$ has i.i.d. standard Gaussian entries, $\boldsymbol{g}' \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_N)$ is independent of $\boldsymbol{G}$. Note that $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$, $\boldsymbol{\theta}_*$, and $\boldsymbol{D}_{\boldsymbol{w}}$ depend on the pretraining task vectors $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M\}$ sampled from $\mathcal{D}_{\text{task}}^{\text{train}}$, while $\boldsymbol{G}$ and $\boldsymbol{g}'$ are independent of these.

We derive an asymptotic characterization for the empirical distribution of the solution to (Primal) via the convex Gaussian min-max theorem (CGMT) [Stojnic, 2013, Thrampoulidis et al., 2015], which has been applied to analyze minimum-norm linear Gaussian problems in recent works [Deng et al., 2019, Montanari et al., 2019, Chang et al., 2021], albeit requiring problem-specific analysis in each case. We outline the key steps below and defer remaining derivations to Appendix C, D.

Our starting point is what is the following Auxiliary Optimization (AO):

$$\min_{\boldsymbol{a}} \max_{\boldsymbol{\lambda}} \ \|\boldsymbol{a}\|\boldsymbol{\lambda}^\top \mathbf{h} + \|\boldsymbol{\lambda}\|\boldsymbol{a}^\top \boldsymbol{g} - \boldsymbol{\lambda}^\top \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}' + \tfrac{1}{2}\|\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2 \tag{5}$$

where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_{p^2})$ and $\mathbf{h} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_N)$ are independent of each other and of all other quantities. The CGMT guarantees that the optimal cost of Eq. (5) converges to the same asymptotic limit as that of (Primal), which sets the stage for establishing that the empirical distribution of the AO solution converges to the same limit as that of the primal [Montanari et al., 2019, Chang et al., 2021].

Solving the AO. Following standard CGMT machinery, we reduce the high-dimensional AO in Eq. (5) to a two-dimensional min-max optimization:

$$\min_{u>0} \max_{\beta\geq 0} \ \frac{\beta u}{2} + \beta\frac{\sum_{i=1}^n \sigma_{n_i}^2}{2u} + \frac{\|\boldsymbol{\theta}_\star\|^2}{2} - \frac{1}{2}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\theta}_\star)^\top \boldsymbol{M}^{-1}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\theta}_\star), \tag{AO}$$

where we define the shorthand $\boldsymbol{M} := \boldsymbol{M}(\beta, u) := \frac{\beta N}{u}\mathbb{I}_{p^2} + \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1}$. Using strict convexity-concavity of the (AO) objective in (u, β), it easily follows that the saddle point (u_*, β_*) is unique whenever $\beta_* > 0$. Moreover, the same reduction that yields (AO) also shows the following closed form for $\boldsymbol{a}_*$, the minimizer of Eq. (5):

$$\boldsymbol{a}_* = \boldsymbol{a}_*(\beta_*, u_*) = \boldsymbol{M}(\beta_*, u_*)^{-1}\big(\beta_* \boldsymbol{g} - \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\theta}_\star\big), \tag{6}$$

where (β_*, u_*) is the unique saddle point of (AO) (when $\beta_* > 0$). Conditional on this saddle point, $\boldsymbol{a}_*$ is Gaussian, with mean and covariance determined by (β_*, τ_*), which as we show next converge to deterministic values in the high dimensional limit.

Deterministic reduction. In the regime specified in Eq. (3), the AO objective converges to a deterministic function of the scalar parameters (u, β). To state this limit compactly, we introduce deterministic quantities defined as the in-probability limits of the following random terms:

$$c_\infty := \mathrm{plim}_{d\to\infty} \frac{1}{d^2}\sum_{i=1}^{N} \sigma_{n_i}^2, \quad s_\infty(\bar{u}, \bar{\beta}) := \mathrm{plim}_{d\to\infty} \frac{1}{d^2}\,\mathrm{tr}\big(\overline{\boldsymbol{M}}^{-1}\big), \tag{7a}$$

$$v_\infty(\bar{u}, \bar{\beta}) := \mathrm{plim}_{d\to\infty} \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star, \tag{7b}$$

where $\bar{\beta}, \bar{u}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}, \overline{\boldsymbol{\theta}}_\star, \overline{\boldsymbol{M}}$ are the normalized versions of $\beta, u, \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}, \boldsymbol{\theta}_\star, \boldsymbol{M}$, chosen so that these quantities remain $O(1)$ (see App. C.1). We prove these convergences (with respect to the randomness in $\boldsymbol{g}$ and the task vectors $\{\boldsymbol{w}_i\}_{i\in[N]}$) in App. D. With these definitions, the AO objective admits the following deterministic limit $D(\bar{u}, \bar{\beta}) := \frac{\bar{\beta}\bar{u}}{2} + \frac{\bar{\beta}}{2\bar{u}}c_\infty - \frac{\bar{\beta}^2}{2}s_\infty(\bar{u}, \bar{\beta}) - \frac{1}{2}v_\infty(\bar{u}, \bar{\beta})$. Since D is strictly convex-concave on $\{\bar{u} > 0, \bar{\beta} > 0\}$, the limiting saddle point $(\bar{\beta}_*, \bar{u}_*)$ is unique and characterized by first-order optimality conditions, which, after algebraic simplification, yield:

$$\begin{aligned} \bar{u}_* &= \frac{\bar{\beta}_*\nu}{a}, \quad \text{where} \quad \Big(1 - \tfrac{\nu}{\rho^2}\Big)\, a = m_a, \\ \bar{\beta}_*^2 &= \frac{a^2\big(\sigma_n^2 + 1 - \rho + \rho c\big) + \rho(1 - a^2c^2)m_a - \rho a z_a^2 m_a'}{\rho^2\left(1 - \frac{\nu}{\rho^2} - \frac{m_a'}{a^2}\right)}. \end{aligned} \tag{8}$$

Here, $m_a := m_\gamma(z_a)$ is the Stieltjes transform of the Marchenko-Pastur law with $\gamma := p/M = \rho/\mu$ (using the scaling in Eq. (3)) evaluated at point $z_a := -1/a - c$. See Definition 4 in the Appendix. In addition, $m_a' := m_\gamma'(z_a)$ denotes the derivative of $m_\gamma(z)$ evaluated at z_a, and $c := \frac{\sigma_n^2+1}{\tau}$.

The appearance of the Marchenko–Pastur Stieltjes transform follows from the spectral structure of $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$. Its explicit form (Eq. (9) in the App.) shows that $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$ can be expressed in terms of blocks of the form $\mathbb{I}_p + \zeta\,\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{w}_i[\mathcal{S}]\boldsymbol{w}_i[\mathcal{S}]^\top$ (for a scalar ζ independent of $\{\boldsymbol{w}_i\}$). Because $\boldsymbol{w}_i[\mathcal{S}]$ are Gaussian, the matrix $\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{w}_i[\mathcal{S}]\boldsymbol{w}_i[\mathcal{S}]^\top$ is Wishart, whose empirical spectral distribution converges to the Marchenko–Pastur law in the proportional limit. Consequently, trace terms such as $\frac{1}{p}\,\mathrm{tr}\big(\mathbb{I}_p + \zeta\,\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{w}_i[\mathcal{S}]\boldsymbol{w}_i[\mathcal{S}]^\top\big)^{-1}$ converge to quantities given in terms of the Stieljes transform.

Evaluation of risk. To evaluate the asymptotic limit of the ID and OOD risks, we first express the per-task loss (defined in Section 2) for fixed task vector $\boldsymbol{w}$ in terms of the error vector $\boldsymbol{a}$ (see Lemma 1 for the expression and proof). Next, we substitute the AO's prediction for the error vector $\boldsymbol{a}$ (Eq. (6)), normalized and evaluated at the deterministic optimal scalars $(\bar{\beta}_*, \bar{u}_*)$: $\boldsymbol{a} = \frac{1}{d}\overline{\boldsymbol{M}}(\bar{\beta}_*, \bar{u}_*)^{-1}\left(\bar{\beta}_*\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star\right)$, where $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_{p^2})$ is independent of $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}, \overline{\boldsymbol{\theta}}_\star, \{\sigma_{n_i}\}_{i=1}^N$. Finally, we derive closed-form limits for ID and OOD losses (Eq. (2)) by taking expectations of the per-task loss over $\boldsymbol{w}$ sampled from $\mathcal{D}_{\text{task}}^{\text{train}}$ and $\mathcal{D}_{\text{task}}$, respectively. See Lemmas 4 and 5 for OOD loss limits, and Lemmas 6 and 7 for ID loss limits.

6 Conclusion

We introduced *in-context benign overfitting*—a task-level analogue of classical benign overfitting, and provided sharp asymptotic characterizations showing that it arises in a minimal feature-selection

model of in-context linear regression. We view our work in a similar light to classical works on benign overfitting in supervised learning setups [Hastie et al., 2019, Belkin et al., 2019], that provide conceptual understanding of phenomena observed in real-world models using minimal models that are amenable to theoretical analysis even if they don't fully capture all properties of real-world data. Moving forward, rigorously establishing the Gaussian equivalence principle in our setting and extending the analysis to richer architectures (e.g., softmax attention) and other task families (e.g., Markov chains) remain natural next steps. Additionally, extension to general task and data covariances beyond isotropic remains an interesting question. As a starting point, in App. H, we conduct experiments with diagonal power-law covariances and observe a stronger form of in-context benign overfitting. More broadly, our results suggest a reassuring message: scaling up need not force a tradeoff between solving known tasks and generalizing to new ones, even though prior theoretical models of ICL might suggest otherwise [Lu et al., 2025, Raventos et al., 2023, Park et al., 2025].

Acknowledgements

This work is supported by a UBC 4YF Fellowship, NSERC Discovery Grant No. 2021- 03677, Alliance Grant ALLRP 581098-22 and a gift from Google.org. Computational resources were also provided by the UBC Advanced Research Computing (ARC) Sockeye cluster.

References

Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. In *Advances in Neural Information Processing Systems*, pages 12372–12382, 2019.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Danil Akhtiamov, Reza Ghane, and Babak Hassibi. One-bit quantization for random features models. *arXiv preprint arXiv:2510.16250*, 2025.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *Int. Conference on Learning Representations*, 2023.

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

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, January 2020. ISSN 2577-0187. doi: 10.1137/20m1336072.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4132–4179. PMLR, 2023.

Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers. *arXiv preprint arXiv:2501.17745*, 2025.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6974–6983, 2021.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024a.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *arXiv preprint arXiv:2409.10559*, 2024b.

Yatin Dandi, Ludovic Stephan, Florent Krzakala, Bruno Loureiro, and Lenka Zdeborová. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36:54754–54768, 2023.

Francesco D'Angelo, Francesco Croce, and Nicolas Flammarion. Selective induction heads: How transformers select causal structures in context. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Puneesh Deora, Bhavya Vasudeva, Tina Behnia, and Christos Thrampoulidis. In-context occam's razor: How transformers prefer simpler hypotheses on the fly. In *Second Conference on Language Modeling*, 2025.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, Eran Malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Spencer Frei and Gal Vardi. Trained transformer classifiers generalize and exhibit benign overfitting in-context. *arXiv preprint arXiv:2410.01774*, 2024.

Deqing Fu, Tian qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

Reza Ghane, Danil Akhtiamov, and Babak Hassibi. Universality in transfer learning for linear models. *Advances in Neural Information Processing Systems*, 37:125729–125779, 2024.

Reza Ghane, Anthony Bao, Danil Akhtiamov, and Babak Hassibi. Gaussian universality for diffusion models. *IEEE Signal Processing Letters*, 33:116–120, 2025.

Sebastian Goldt, Galen Reeves, Marc Mézard, Florent Krzakala, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with two-layer neural networks. *arXiv preprint arXiv:2006.14709*, 2020.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *arXiv preprint arXiv:2009.07669*, 2020.

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

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Tengyuan Liang and Pragya Sur. A precise high-dimensional asymptotic theory for boosting and min-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

Bruno Loureiro, Gabriele Sicuro, Cédric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborová. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021.

Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2025.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

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

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

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

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

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

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics.

Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. *Advances in Neural Information Processing Systems*, 30, 2017.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? the extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, pages 27680–27708. PMLR, 2023.

Nived Rajaraman, Marco Bondaschi, Ashok Vardhan Makkuva, Kannan Ramchandran, and Michael Gastpar. Transformers on markov data: Constant depth suffices. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *arXiv preprint arXiv:2311.08360*, 2023.

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

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society, 2023.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 1683–1709. PMLR, 2015.

Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. Precise error analysis of regularized m -estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018. doi: 10.1109/TIT.2018.2840720.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. *arXiv preprint arXiv:2212.07677*, 2022.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Garrett G Wen, Hong Hu, Yue M Lu, Zhou Fan, and Theodor Misiakiewicz. When does gaussian equivalence fail and how to fix it: Non-universal behavior of random features with quadratic scaling. *arXiv preprint arXiv:2512.03325*, 2025.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter L Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Ruiqi Zhang, Jingfeng Wu, and Peter Bartlett. In-context learning of a linear transformer block: Benefits of the mlp component and one-step gd initialization. *Advances in Neural Information Processing Systems*, 37:18310–18361, 2024.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

Appendix

A Related Work **15**

B Gaussian Equivalent Data Model **16**

C Forming the Auxiliary Optimization Problem **18**

C.1 Solving the Auxiliary Optimization. 19

C.2 Simplifying Per-Task Loss 19

D Proofs for Min Norm AO **20**

D.1 Wishart Matrix Asymptotics 20

D.2 General Notation 20

D.3 Asymptotic limit of $S(\tau, \beta)$ 21

D.4 Asymptotic limit of the quadratic forms 22

D.5 First Order Optimality Conditions of $D(\bar{u}, \bar{\beta})$ 23

E Proof for LS AO **26**

F Asymptotic Risk **27**

F.1 Asymptotic Risk of Minimum-Norm LGP 27

F.2 Asymptotic Risk of LS LGP 35

G Helper Lemmas **37**

H Additional Results and Details of Experimental Settings **42**

A Related Work

To deconstruct the mechanics of in-context learning, a growing body of literature has turned to controlled synthetic environments where Transformers are pre-trained from scratch on canonical function classes. Linear regression has served as a primary testbed for these investigations [Garg et al., 2022, Raventos et al., 2023, Akyürek et al., 2023, von Oswald et al., 2022, Ahn et al., 2023, Zhang et al., 2025], alongside sequence modeling tasks defined by Markov chains [Park et al., 2025, Edelman et al., 2024, Rajaraman et al., 2024, D'Angelo et al., 2025, Deora et al., 2025]. A central theme in this line of inquiry is algorithmic discovery: determining whether trained Transformers implement known estimation procedures, such as gradient descent variants for regression [von Oswald et al., 2022, Ahn et al., 2023, Fu et al., 2024] or induction-head mechanisms for Markov processes [Edelman et al., 2024, Rajaraman et al., 2024, Chen et al., 2024b]. Additionally, a growing body of work also investigates the training dynamics of in-context learning for linear regression, specifically examining the optimization dynamics for both one-layer linear attention [Zhang et al., 2025, 2023, 2024] and softmax attention [Chen et al., 2024a].

Retreival and Learning Modes of ICL. Most relevant to our work are studies investigating the dual nature of ICL: the retrieval and learning modes [Pan et al., 2023]. Empirical works involving transformers trained on finite task sets have identified task diversity thresholds—critical points where the model transitions from retrieval (good ID generalization) to learning (good OOD tasks) [Raventos et al., 2023, Park et al., 2025]. While Wu et al. [2023] provided an initial theoretical quantification of this transition in linear attention for in-context linear regression, our analysis is most closely related to and partly inspired by Lu et al. [2025]. They derive precise risk asymptotics in a proportional

limit similar to ours (Eq. (3)), but with a crucial distinction: their setup is restricted to $\rho = 1$ (utilizing all features). This difference is fundamental; as we demonstrate in Section 4 (Fig. 3), fixing $\rho = 1$ couples model scale with the ambient dimension. Under this constraint, the in-context benign overfitting phenomenon disappears, highlighting the necessity of the feature-selection mechanism to decouple model capacity from data dimensionality.

Finally, we note that Frei and Vardi [2024] also coin the term *in-context benign overfitting* though in a fundamentally different context. Their work focuses on in-context binary classification and interprets *overfitting* through the lens of label noise: they show that models can fit prompts with flipped labels (noise) while still generalizing to the query. This is strictly analogous to classical benign overfitting over samples. In contrast, our framework applies the concept to the task level, describing models that overfit (memorize) the pre-training *task* distribution while generalizing to novel distributions.

Gaussian Universality and Applications of CGMT. The principle that Gaussian systems matching a non-Gaussian system to first and second order are predictive of properties of the latter is broadly known as *universality*. Its roots lie in classical random matrix theory results on eigendistributions of random ensembles [Tao, 2023]. The specific form relevant here, where the "system" refers to a random optimization problem and "properties" to quantities such as its optimal cost or empirical distribution of its solution, has been developed within the high-dimensional statistics literature, from early applications in compressed sensing (e.g. [Montanari and Nguyen, 2017, Oymak and Tropp, 2018, Panahi and Hassibi, 2017, Abbasi et al., 2019]) to more recent machine learning settings (e.g. [Hastie et al., 2019, Mei and Montanari, 2019, Goldt et al., 2020, Bosch et al., 2023, Montanari and Saeed, 2022, Hu and Lu, 2022, Ghane et al., 2024, Akhtiamov et al., 2025]). The terminology *Gaussian equivalence principle* was introduced by Mei and Montanari [2019], Goldt et al. [2020], who applied it to study linear regression with random features. Since then, numerous extensions have applied and confirmed the gaussian universality principle in various contexts, e.g., [Ghane et al., 2024, Dandi et al., 2023, Gerace et al., 2024, Pesce et al., 2023, Misiakiewicz and Saeed, 2024, Ghane et al., 2025]. It is also understood that universality does not hold always and very recent work has extended the methodology to handle such cases [Wen et al., 2025]. Here, we identify another instance where the Gaussian equivalence principle applies. Fig. 2 empirically demonstrates this. Formalizing this equivalence rigorously is left for future work, as our focus here is on deriving predictions and establishing in-context benign overfitting; a formal proof could likely build on recent formal universality equivalence proofs such as [Hu and Lu, 2022, Montanari et al., 2023].

In order to obtain deterministic limits for the ID and OOD risks presented in Theorem 1, 2, we use the CGMT framework [Stojnic, 2013, Thrampoulidis et al., 2015]. This framework allows us to derive an asymptotic characterization for the empirical distribution of the solution to (Primal). The application of the CGMT to analyze minimum-norm linear Gaussian problems (Definition 3) appears in various recent works, e.g., [Deng et al., 2019, Montanari et al., 2019, Loureiro et al., 2021, Chang et al., 2021, Liang and Sur, 2020]. While the CGMT framework provides a general recipe, its application still requires problem-specific analysis, which we carry out in our work since, to the best of our knowledge, no plug-and-play result exists for our specific setting.

B Gaussian Equivalent Data Model

Below, we define the Linear Gaussian problems (LGPs) that are central to our analyis.

Definition 3 (Linear Gaussian MSE Problem)**.** *An* $\mathrm{LGP}(\boldsymbol{\theta}_*, \boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_\epsilon)$ *with* $\boldsymbol{\Sigma}_g, \boldsymbol{\Sigma}_\epsilon$ *positive definite, is a linear regression problem with squared loss on* N *samples* $\{(\boldsymbol{g}_i, y_i)\}_{i=1}^N$ *generated as follows: features* $\boldsymbol{g}_i \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_{\tilde{p}})$ *and labels* $y_i = \boldsymbol{g}_i^\top \boldsymbol{\Sigma}_g^{1/2} \boldsymbol{\theta}_* + \epsilon_i$*, where* $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$*. Let* $\tilde{p}$ *denote the dimension of the regressor* $\boldsymbol{\theta}_* \in \mathbb{R}^{\tilde{p}}$*, and let* $\boldsymbol{G} = [\boldsymbol{g}_1, \dots, \boldsymbol{g}_N]^\top \in \mathbb{R}^{N \times \tilde{p}}$ *be the feature matrix. The estimator is defined as*

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_g^{-1/2} \hat{\boldsymbol{a}} + \boldsymbol{\theta}_* \,,$$

with the error vector $\hat{\boldsymbol{a}}$ *given as follows.*

- *If* $\tilde{p} < N$*: the least-squares solution* $\hat{\boldsymbol{a}} = \boldsymbol{a}_{\mathrm{LS}} := \arg\min_{\boldsymbol{a}} \|\boldsymbol{\epsilon} - \boldsymbol{G}\boldsymbol{a}\|^2$.
- *If* $\tilde{p} > N$*: the minimum-norm interpolator* $\hat{\boldsymbol{a}} = \boldsymbol{a}_{\mathrm{MN}} := \arg\min_{\boldsymbol{a}} \|\boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_*\| \quad s.t. \quad \boldsymbol{G}\boldsymbol{a} = \boldsymbol{\epsilon}$.

Next, we discuss the moment-matching process for the first and second moments involving features $\boldsymbol{g}$ and labels y in Definition 2 with those from the original data model discussed in Section 2. Recall that for the Gaussian equivalent data model, the features are i.i.d $\boldsymbol{g}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}})$ for all $i \in [N]$, where $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$ matches the second moment of the original feature matrix $\boldsymbol{H}_{\mathcal{S}_p}$ by definition as $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} := \frac{1}{N}\mathbb{E}\Big[\boldsymbol{H}_{\mathcal{S}_p}^\top \boldsymbol{H}_{\mathcal{S}_p}\Big]$. Below, we show that the first moments also match, *i.e.*, $\mathbb{E}[\boldsymbol{g}] = \mathbb{E}[\mathbf{h}_{\mathcal{S}_p}(X)]$.

We compute the first and second moments of feature matrix $\boldsymbol{H}_{\mathcal{S}_p}$ by finding the expected value of each feature row. By definition, we have

$$\mathbb{E}[\mathbf{h}_{\mathcal{S}_p}(X)] = \mathbb{E}[\texttt{vec}(\boldsymbol{x}_q[\mathcal{S}]\hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}])^\top] = \mathbb{E}[\hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}]] \otimes \mathbb{E}[\boldsymbol{x}_q[\mathcal{S}]] = 0,$$

where the second equality uses the independence of the query vector from the the rest of the context, and the third equality uses $\boldsymbol{x}_q \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I}_d)$.

Next, we compute the second moment of $\boldsymbol{H}_{\mathcal{S}_p}$. We have

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} &:= \frac{1}{n}\mathbb{E}\left[\boldsymbol{H}_{\mathcal{S}_p}^\top \boldsymbol{H}_{\mathcal{S}_p}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\mathbf{h}_i\mathbf{h}_i^\top] \\ &= \frac{\sigma_x^4}{TM}\sum_{i=1}^{M}\left[(\sigma_x^2\|\boldsymbol{w}_i\|^2 + \sigma_n^2)\mathbb{I}_{p^2} + \sigma_x^2(T+1)(\boldsymbol{w}_i[\mathcal{S}]\boldsymbol{w}_i[\mathcal{S}]^\top) \otimes \mathbb{I}_p\right]. \end{aligned} \tag{9}$$

Here, the first equality uses the independence of the rows of $\boldsymbol{H}_{\mathcal{S}_p}$, and the second equality follows by using Lemma 16.

Recall that the labels in the Gaussian equivalent problem are generated as $y_i = \boldsymbol{g}_i^\top \boldsymbol{\theta}_\star + \epsilon_i$, where we select $\boldsymbol{\theta}_\star$ and noise ϵ_i such that the moments, $\mathbb{E}[y]$, $\boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}} := \mathbb{E}[\boldsymbol{H}_{\mathcal{S}_p}^\top \boldsymbol{y}]$ and $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$ also match with that of the original problem.

First, in the original model, $\mathbb{E}[y] = 0$. In the Gaussian equivalent data model, $\mathbb{E}[y] = \mathbb{E}[\boldsymbol{g}]^\top \boldsymbol{\theta}_\star + \mathbb{E}[\epsilon] = \mathbb{E}[\epsilon]$. To match, we require $\mathbb{E}[\epsilon] = 0$.

Next, as the labels y are independent, $\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top]$ is a diagonal matrix, and we only need to look at the entries $\mathbb{E}[y^2]$.

In the original model, we have

$$\mathbb{E}[y^2] = \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x} + \epsilon)^2] = \sigma_x^2\|\boldsymbol{w}\|^2 + \sigma_n^2.$$

For the Gaussian equivalent data model,

$$\mathbb{E}[y^2] = \boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} \boldsymbol{\theta}_\star + \mathbb{E}[\epsilon^2].$$

Therefore, for the Gaussian equivalent data model $\epsilon_i \sim \mathcal{N}(0, \sigma_{n_i}^2)$, where

$$\sigma_{n_i}^2 = \sigma_x^2\|\boldsymbol{w}_i\|^2 + \sigma_n^2 - \boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} \boldsymbol{\theta}_\star =: d_i^2. \tag{10}$$

Here, $\boldsymbol{w}_i$ corresponds to the task vector for sample $i \in [n]$.

Recall that $\boldsymbol{G} \in \mathbb{R}^{N \times p^2}$ has i.i.d. entries $G_{ij} \sim \mathcal{N}(0,1)$, so that $\boldsymbol{G}\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{1/2}$ has i.i.d. rows distributed as $\mathcal{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}})$. Therefore, the remaining second moment-matching condition gives

$$\boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}} = \mathbb{E}[\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{1/2} \boldsymbol{G}^\top \boldsymbol{G} \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{1/2} \boldsymbol{\theta}_\star] = \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{1/2} \mathbb{E}[\boldsymbol{G}^\top \boldsymbol{G}] \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{1/2} \boldsymbol{\theta}_\star = N \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} \boldsymbol{\theta}_\star,$$

which implies

$$\boldsymbol{\theta}_\star = \frac{1}{N} \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}}. \tag{11}$$

Here, using Lemma 16, we have

$$\begin{aligned}\boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}} &= \mathbb{E}\left[\sum_{i=1}^{n} \mathbf{h}_i y_i\right] = \sum_{i=1}^{M} \frac{n}{M}\mathbb{E}[\mathbf{h}_i y_i] \\ &= \frac{n}{M}\sum_{i=1}^{M} \sigma_x^4 \boldsymbol{w}_i[\mathcal{S}] \otimes \boldsymbol{w}_i[\mathcal{S}]. \end{aligned} \tag{12}$$

C Forming the Auxiliary Optimization Problem

Min norm using Gaussian equivalent. First, recall the minimum norm defined using the Gaussian equivalent data. We have

$$\min_{\boldsymbol{\theta}} \ \tfrac{1}{2}\|\boldsymbol{\theta}\|^2 \ \text{ subject to } \ \boldsymbol{G}(\boldsymbol{\theta}-\boldsymbol{\theta}_\star) = \boldsymbol{D_w}\boldsymbol{g}', \tag{13}$$

where $\boldsymbol{D_w}$ is a diagonal matrix with the ith entry equal to σ_{n_i} for sample $i \in [n]$. Lastly, we use $\boldsymbol{g}' \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I})$ to denote noise that is independent of entries in $\boldsymbol{G} \in \mathbb{R}^{N\times p^2}$. Here, the rows of $\boldsymbol{G}$, $\boldsymbol{g}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}})$ are i.i.d sampled for all $i \in [N]$.

Changing variable

$$\boldsymbol{\Sigma}^{1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}(\boldsymbol{\theta}-\boldsymbol{\theta}_\star) =: \boldsymbol{a}, \tag{14}$$

and substituting $\boldsymbol{G} \leftarrow \boldsymbol{G}\boldsymbol{\Sigma}^{1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}$, where $\boldsymbol{G}$ now refers to entries from unit variance isotropic Gaussian, Eq. (13) is now

$$\min_{\boldsymbol{a}} \ \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2 \quad \text{subject to} \quad \boldsymbol{G}\boldsymbol{a} = \boldsymbol{D_w}\boldsymbol{g}'.$$

Using Lagrangian formulation, the solution to the above problem is the same as the one to the following unconstrained min-max problem.

$$\min_{\boldsymbol{a}}\max_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^\top \boldsymbol{G}\boldsymbol{a} - \boldsymbol{\lambda}^\top \boldsymbol{D_w}\boldsymbol{g}' + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2.$$

Applying the CGMT Thrampoulidis et al. [2015, 2018], the corresponding Auxiliary Optimization (AO) is

$$\min_{\boldsymbol{a}}\max_{\boldsymbol{\lambda}} \ \|\boldsymbol{a}\|\boldsymbol{\lambda}^\top \mathbf{h} + \|\boldsymbol{\lambda}\|\boldsymbol{a}^\top \boldsymbol{g} - \boldsymbol{\lambda}^\top \boldsymbol{D_w}\boldsymbol{g}' + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2$$

Let $\beta = \|\boldsymbol{\lambda}\|$, and $\hat{\boldsymbol{\lambda}}$, be the corresponding unit norm direction, then the above can be written as

$$\begin{aligned}
&\min_{\boldsymbol{a}} \max_{\hat{\boldsymbol{\lambda}},\beta\geq 0} \beta\hat{\boldsymbol{\lambda}}^\top(\|\boldsymbol{a}\|\mathbf{h} - \boldsymbol{D_w}\boldsymbol{g}') + \beta\boldsymbol{a}^\top\boldsymbol{g} + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2 \\
&= \min_{\boldsymbol{a}} \max_{\beta\geq 0} \beta\|\ (\|\boldsymbol{a}\|\mathbf{h} - \boldsymbol{D_w}\boldsymbol{g}')\| + \beta\boldsymbol{a}^\top\boldsymbol{g} + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2,
\end{aligned}$$

where the final step uses the fact that the objective maximizes over $\hat{\boldsymbol{\lambda}}$ when $\hat{\boldsymbol{\lambda}} = \frac{\|\boldsymbol{a}\|\mathbf{h}-\boldsymbol{D_w}\boldsymbol{g}'}{\|\|\boldsymbol{a}\|\mathbf{h}-\boldsymbol{D_w}\boldsymbol{g}'\|}$.

The objective is convex in $\boldsymbol{a}$ and linear, hence concave, in β. As shown in previous works, e.g. [Chang et al., 2021], the constraint sets can be restricted to compact subsets, hence using Sion's minimax theorem we can swap the order of the optimization to min-max to $\max_\beta \min_{\boldsymbol{a}}$:

$$\max_{\beta\geq 0}\min_{\boldsymbol{a}} \beta\|\ (\|\boldsymbol{a}\|\mathbf{h} - \boldsymbol{D_w}\boldsymbol{g}')\| + \beta\boldsymbol{a}^\top\boldsymbol{g} + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2$$

Next, using the variational characterization of the sqrt function (again, see for example [Chang et al., 2021, Sec. B.4]), the above optimization is

$$\max_{\beta\geq 0}\min_{\boldsymbol{a},u>0} \tfrac{\beta u}{2} + \beta\tfrac{(\|\boldsymbol{a}\|\mathbf{h}-\boldsymbol{D_w}\boldsymbol{g}')^2}{2u} + \beta\boldsymbol{a}^\top\boldsymbol{g} + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2 \tag{15}$$

Next, as $n \to \infty$, we have $\frac{\|\mathbf{h}\|^2}{n} \to 1$ and $\frac{\mathbf{h}^\top\boldsymbol{D_w}\boldsymbol{g}'}{n} \to 0$, and $\frac{\|\boldsymbol{D_w}\boldsymbol{g}'\|^2}{n} - \frac{1}{n}\sum_{i=1}^n d_i^2 \to 0$ in probability. Here and onwards, we denote $d_i = \sigma_{n_i}$. Replacing these random quantities in Eq. (15) yields the following asymptotically equivalent optimization wpa1:

$$\max_{\beta\geq 0}\min_{\boldsymbol{a},u>0} \tfrac{\beta u}{2} + \beta\tfrac{n\|\boldsymbol{a}\|^2+\sum_{i=1}^n d_i^2}{2u} + \beta\boldsymbol{a}^\top\boldsymbol{g} + \tfrac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a} + \boldsymbol{\theta}_\star\|^2 \tag{16}$$

Notice, we have a quadratic in $\boldsymbol{a}$. Minimizing the above w.r.t. $\boldsymbol{a}$, the optimal is at

$$\boldsymbol{a}_* := \boldsymbol{a}_*(\beta, u) = \boldsymbol{M}^{-1}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\theta}_\star), \quad \boldsymbol{M} := \tfrac{\beta n}{u}\mathbb{I} + \boldsymbol{\Sigma}^{-1}_{\boldsymbol{H}_{\mathcal{S}_p}}. \tag{17}$$

Plugging this back in Eq. (16), it simplifies to

$$\max_{\beta\geq 0}\min_{u>0} \ \underbrace{\tfrac{\beta u}{2} + \beta\tfrac{\sum_{i=1}^n d_i^2}{2u} - \tfrac{1}{2}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\theta}_\star)^\top\boldsymbol{M}^{-1}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\theta}_\star) + \tfrac{\|\boldsymbol{\theta}_\star\|^2}{2}}_{R(u,\beta)} \tag{AO}$$

C.1 Solving the Auxiliary Optimization.

The subsequent analysis will use the saddle point $(\bar{u}_\star, \bar{\beta}_\star)$ of the deterministic limit objective defined below. Formally relating $(\bar{u}_\star, \bar{\beta}_\star)$ to the finite-dimensional saddle point requires a uniform convergence argument for the AO objective for which we refer the reader to Chang et al. [2021, Sec. B.4]. We treat this as a standard technical condition and proceed.

Choice of scaling. We next identify the natural normalization under the asymptotic regime. For our setup, note that $\|\boldsymbol{\theta}_\star\| = \Theta(d^{3/2})$, $\lambda_i = \Theta(1/d^3)$ for all $i \in [p^2]$, where λ_i denote the Eigen values of $\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}$. Moreover, $\tilde{\lambda}_i = \Theta(d^3)$, where $\tilde{\lambda}_i$ denote the Eigen values of $\mathbf{M}$. Using this, and Eqs. (3), (9), (11), (12) and (17), we define the following rescaled quantities:

$$\bar{u} := u/d, \quad \bar{\beta} := \beta/d^2, \tag{18}$$

$$\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}} := d^3 \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} = \frac{1}{\tau}\left(\sigma_n^2 + \frac{1}{M}\sum_{i=1}^{M}\frac{\|\boldsymbol{w}_i\|^2}{d}\right)\mathbb{I}_{p^2} + \frac{1}{M}\sum_{i=1}^{M}(\boldsymbol{w}_i[\mathcal{S}]\boldsymbol{w}_i[\mathcal{S}]^\top) \otimes \mathbb{I}_p, \tag{19}$$

$$\overline{\boldsymbol{M}} := \tfrac{\boldsymbol{M}}{d^3} = \frac{\bar{\beta}\nu}{\bar{u}}\mathbb{I} + \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{H}_{\mathcal{S}_p}} \tag{20}$$

$$\overline{\boldsymbol{\theta}}_\star := \frac{\boldsymbol{\theta}_\star}{d} = \frac{1}{dn}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p}\boldsymbol{y}} = \overline{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{H}_{\mathcal{S}_p}}\left(\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{w}_i[\mathcal{S}] \otimes \boldsymbol{w}_i[\mathcal{S}]\right). \tag{21}$$

Next, we normalize the auxiliary objective (AO) by d^3, and get the following objective

$$\begin{aligned} D_d(u,\beta) &:= \frac{1}{d^3}R(u,\beta) \\ &= \bar{\beta}\bar{u} + \frac{\bar{\beta}}{2\bar{u}}\frac{1}{d^2}\sum_{i=1}^{n} d_i^2 - \frac{1}{2d^2}\bar{\beta}^2\boldsymbol{g}^\top\overline{\boldsymbol{M}}^{-1}\boldsymbol{g} - \frac{1}{2d}\overline{\boldsymbol{\theta}}_\star^\top\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star + \frac{1}{d^{7/2}}\boldsymbol{g}^\top\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star + \frac{\|\overline{\boldsymbol{\theta}}_\star\|^2}{2d}. \end{aligned}$$

Using Lemma 14, it follows that

$$\lim_{d\to\infty}\frac{1}{2d^2}\boldsymbol{g}^\top\overline{\boldsymbol{M}}^{-1}\boldsymbol{g} = \lim_{d\to\infty}\frac{1}{d^2}\operatorname{tr}(\overline{\boldsymbol{M}}^{-1}) =: s_\infty(\bar{u},\bar{\beta}), \qquad \lim_{d\to\infty}\frac{1}{d^{7/2}}\boldsymbol{g}^\top\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star = 0. \tag{22}$$

Further, define the deterministic limits

$$c_\infty := \lim_{d\to\infty}\frac{1}{d^2}\sum_{i=1}^{n} d_i^2, \quad v_\infty(\bar{u},\bar{\beta}) := \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star. \tag{23}$$

Using Eq. (22), Eq. (23) we have the following deterministic limit of $D_d(\tau,\beta)$

$$D(\bar{u},\bar{\beta}) = \tfrac{\bar{\beta}\bar{u}}{2} + \tfrac{\bar{\beta}}{2\bar{u}}c_\infty - \tfrac{\bar{\beta}^2}{2}s_\infty(\bar{u},\bar{\beta}) - \tfrac{1}{2}v_\infty(\bar{u},\bar{\beta}). \tag{24}$$

Here, we first compute $\lim_{d\to\infty} D_d(u,\beta)$, and then, since $\lim_{d\to\infty}\frac{\|\overline{\boldsymbol{\theta}}_\star\|^2}{2d} = 0.5$, we omit it as it is independent of $\bar{\beta}, \bar{u}$.

C.2 Simplifying Per-Task Loss

Lemma 1. *The per-task loss evaluated at fixed $\boldsymbol{w}$ with respect to the error vector $\boldsymbol{a}$ is given by.*

$$\begin{aligned} \mathcal{L}(\boldsymbol{a};\boldsymbol{w}) = \underbrace{\boldsymbol{a}^\top\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{a}}_{\textit{Term-1(a)}} + \underbrace{\frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star}_{\textit{Term-1(b)}} + \underbrace{2\,\frac{1}{\sqrt{d}}\boldsymbol{a}^\top\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star}_{\textit{Term-1(c)}} \\ \underbrace{- 2\left(\frac{1}{d^{1/2}}\boldsymbol{a}^\top\overline{\boldsymbol{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}(\boldsymbol{w}\otimes\boldsymbol{w}) + \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top(\boldsymbol{w}\otimes\boldsymbol{w})\right)}_{\textit{Term-2}} + \frac{\|\boldsymbol{w}\|^2}{d} + \sigma_n^2. \end{aligned}$$

Proof.

$$\begin{aligned}\mathcal{L}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}; \boldsymbol{w}) &= \mathbb{E}_{X,\epsilon}[\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbf{h}_{\mathcal{S}_p}(X) - y_q]^2. \\ &= \hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X)\mathbf{h}_{\mathcal{S}_p}(X)^\top]\hat{\boldsymbol{\theta}}_{\mathcal{S}_p} + \mathbb{E}_{X,\epsilon}[y_q^2] - 2\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q] \\ &= \underbrace{\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X)\mathbf{h}_{\mathcal{S}_p}(X)^\top]\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}}_{\text{Term-1}} + \sigma_x^2\|\boldsymbol{w}\|^2 + \sigma_n^2 - 2\underbrace{\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q]}_{\text{Term-2}}. \end{aligned} \tag{25}$$

Let $\boldsymbol{\Sigma}(\boldsymbol{w}) := \mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X)\mathbf{h}_{\mathcal{S}_p}(X)^\top]$. Then, using Lemma 16 and Eq. (3), define

$$\overline{\boldsymbol{\Sigma}}(\boldsymbol{w}) := d^3\boldsymbol{\Sigma}(\boldsymbol{w}) = \frac{1}{\tau}\left(\frac{\|\boldsymbol{w}\|^2}{d} + \sigma_n^2\right)\mathbb{I}_{p^2} + \left(1 + \frac{1}{\tau d}\right)(\boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top) \otimes \mathbb{I}_p.$$

To simplify Term-1 and Term-2, we use $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p} = \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star$ from Eq. (14), and Lemma 16.

First, we have Term-1:

$$\begin{aligned}(\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star)^\top \boldsymbol{\Sigma}(\boldsymbol{w})(\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star) &= \boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}_\star + 2\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}_\star \\ &= \underbrace{\boldsymbol{a}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a}}_{\text{Term-1(a)}} + \underbrace{\frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top \overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star}_{\text{Term-1(b)}} + 2\underbrace{\frac{1}{\sqrt{d}}\boldsymbol{a}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star}_{\text{Term-1(c)}}. \end{aligned} \tag{26}$$

Next, we have Term-2:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top \mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q] &= (\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{a} + \boldsymbol{\theta}_\star)^\top(\sigma_x^4 \boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}]) \\ &= \sigma_x^4 \boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}] + \sigma_x^4 \boldsymbol{\theta}_\star^\top \boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w} \\ &= \frac{1}{d^2}\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{w} \otimes \boldsymbol{w}[\mathcal{S}] + \frac{1}{d^2}\boldsymbol{\theta}_\star^\top \boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}] \\ &= \frac{1}{d^{1/2}}\boldsymbol{a}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}(\boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}]) + \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top(\boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}]). \end{aligned} \tag{27}$$

Substituting Eqs. (26) and (27) in Eq. (25) finishes the proof. □

D Proofs for Min Norm AO

D.1 Wishart Matrix Asymptotics

Definition 4. *Let μ be a probability measure on $\mathbb{R}$. The Stieltjes transform of μ, denoted by $m_\mu(z)$, is a function defined for all complex numbers $z \in \mathbb{C} \setminus \operatorname{supp}(\mu)$ by:*

$$m_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z}\,\mathrm{d}\mu(x).$$

Remark 1 (Stieltjes transform of the Marchenko-Pastur law)**.** *In the asymptotic regime where $p, M \to \infty$ with $p/M \to \gamma$, the limiting spectral distribution of the normalized Wishart matrix $\boldsymbol{W}_p$ is the Marchenko-Pastur law μ_γ. Its Stieltjes transform $m_\gamma(z)$ is determined by the quadratic equation:*

$$z\gamma m_\gamma(z)^2 + (z + \gamma - 1)m_\gamma(z) + 1 = 0.$$

For $z \in \mathbb{C} \setminus \operatorname{supp}(\mu_\gamma)$:

$$m_\gamma(z) = \frac{-(z + \gamma - 1) - \sqrt{(z + \gamma - 1)^2 - 4z\gamma}}{2z\gamma}.$$

D.2 General Notation

We use the following notation throughout the appendix. Let $\boldsymbol{G}$ be a $M \times p$ random matrix with independent and identically distributed entries $G_{ij} \sim \mathcal{N}(0, 1)$. Let $\boldsymbol{g}_i$ denote row i of $\boldsymbol{G}$. Let

$\boldsymbol{W}_p = \frac{1}{M}\boldsymbol{G}^\top \boldsymbol{G}$ denote the normalized Wishart matrix. Let $\lambda_1, \ldots, \lambda_p$ denote its eigenvalues. Let $\boldsymbol{g} \sim \mathcal{N}(0, \mathbb{I}_{p^2})$ be independent of $\boldsymbol{W}_p$, and let $\boldsymbol{w} \sim \mathcal{N}(0, \mathbb{I}_d)$ be independent of $(\boldsymbol{g}, \boldsymbol{W}_p)$.

Let $\otimes$ denote the Kronecker product, and $\mathrm{tr}(\cdot)$ denote the trace. We state some of its properties below. For conformable matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, vectors $\boldsymbol{a}, \boldsymbol{b}$,

$$(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}, \tag{28}$$

$$\mathrm{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A})\,\mathrm{tr}(\boldsymbol{B}), \tag{29}$$

$$(\boldsymbol{A} \otimes \mathbb{I})\mathtt{vec}(\boldsymbol{C}) = \mathtt{vec}(\boldsymbol{A}\boldsymbol{C}), \tag{30}$$

$$(\boldsymbol{A} \otimes \mathbb{I})(\boldsymbol{B} \otimes \mathbb{I}) = (\boldsymbol{A}\boldsymbol{B}) \otimes \mathbb{I}, \tag{31}$$

$$\mathtt{vec}(\boldsymbol{a}\boldsymbol{b}^\top)\mathtt{vec}(\boldsymbol{a}\boldsymbol{b}^\top)^\top = \boldsymbol{b}\boldsymbol{b}^\top \otimes \boldsymbol{a}\boldsymbol{a}^\top. \tag{32}$$

Also,

$$\mathtt{vec}(\boldsymbol{A})^\top \mathtt{vec}(\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{B}). \tag{33}$$

Let $\Pi_{p\times d}$ denote the permutation matrix with all entries 0 or 1, and each row has exactly one 1 and most one 1 per column. Let $a, b, c, b', c', \beta, \kappa \in \mathbb{R}$ be constants. Also, let $\mathcal{S} \subset \{1, \ldots, d\}$ denote a subset of indices, with $|\mathcal{S}| = p$. Next, we define matrices

$$\begin{aligned}
\boldsymbol{A} &= c\mathbb{I}_p + b\boldsymbol{W}_p, \quad \boldsymbol{B} = a\mathbb{I}_p + \boldsymbol{A}^{-1} = a\mathbb{I}_p + (c\mathbb{I}_p + b\boldsymbol{W}_p)^{-1},\\
\boldsymbol{\Sigma} &= \boldsymbol{A} \otimes \mathbb{I}_p = c\mathbb{I}_{p^2} + b\boldsymbol{W}_p \otimes \mathbb{I}_p,\\
\mathbf{M} &= \boldsymbol{B} \otimes \mathbb{I}_p = a\mathbb{I}_{p^2} + \boldsymbol{\Sigma}^{-1},\\
\boldsymbol{\Sigma}(\boldsymbol{w}) &= \left(c'\tfrac{\|\boldsymbol{w}\|^2}{d} + b'\right)\mathbb{I}_{p^2} + (\boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top) \otimes \mathbb{I}_p,\\
\boldsymbol{\Sigma}(\boldsymbol{w}_j) &= \left(c'\tfrac{\|\boldsymbol{w}_j\|^2}{d} + b'\right)\mathbb{I}_{p^2} + (\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top) \otimes \mathbb{I}_p,
\end{aligned}$$

where $\boldsymbol{w}_j = \boldsymbol{g}_j$, the jth row of $\boldsymbol{G}$. Next, define vectors

$$\begin{aligned}
\boldsymbol{v} &= \mathtt{vec}(\boldsymbol{W}_p), \qquad \boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} \boldsymbol{g}_i \otimes \boldsymbol{g}_i\right) = \boldsymbol{\Sigma}^{-1}\boldsymbol{v},\\
\boldsymbol{u} &= \beta\boldsymbol{g} - d^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta},
\end{aligned}$$

and scalars

$$z_a = -\frac{(1+ac)}{ab}, \quad z_c = -\frac{c}{b}.$$

We consider the asymptotic regime where $d, p, M \to \infty$ with ratios

$$\frac{p}{d} =: \rho, \qquad \frac{M}{d} =: \mu,$$

where $\rho, \mu \in (0, \infty)$. The following lemmas are in the asymptotic regime $d, p, M \to \infty$ with fixed ratios, we will use shorthand $\lim_{d\to\infty}$ to denote $\lim_{d,p,M\to\infty}$.

D.3 Asymptotic limit of $S(\tau, \beta)$

Lemma 2. *The normalized trace of the $p^2 \times p^2$ random matrix $\mathbf{M}^{-1}$ converges in probability to the following deterministic limit:*

$$\mathrm{plim}_{d\to\infty} \frac{1}{d^2}\,\mathrm{tr}(\mathbf{M}^{-1}) = \frac{\rho^2}{a}\left(1 - \frac{1}{ab}m_\gamma(z_a)\right).$$

Proof. Using properties of the Kronecker product (Eqs. (28) and (29)), we have:

$$\mathrm{tr}(\mathbf{M}^{-1}) = \mathrm{tr}(\boldsymbol{B}^{-1} \otimes \mathbb{I}_p) = p\,\mathrm{tr}(\boldsymbol{B}^{-1}) = p\,\mathrm{tr}\left(\left(a\mathbb{I}_p + (c\mathbb{I}_p + b\boldsymbol{W}_p)^{-1}\right)^{-1}\right).$$

The normalized trace is:

$$\frac{1}{d^2}\operatorname{tr}(\boldsymbol{M}^{-1}) = \frac{p\rho^2}{p^2}\sum_{i=1}^{p}\left(a + \frac{1}{c+b\lambda_i}\right)^{-1} = \frac{\rho^2}{p}\sum_{i=1}^{p}\frac{b\lambda_i + c}{ab\lambda_i + ac + 1}.$$

As $p, M \to \infty$, and the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ. The sum converges to the integral:

$$\mathcal{I} = \frac{\rho^2}{a}\int \frac{\lambda + c/b}{\lambda + (ac+1)/ab}\,\mathrm{d}\mu_\gamma(\lambda).$$

Using Lemma 10 ($\mathcal{I}_1$ with $z_1 = z_c, z_2 = z_a$) gives the final expression. □

D.4 Asymptotic limit of the quadratic forms

Lemma 3. *Let* $Q_p := \boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}$. *Then we have,*

$$\operatorname{plim}_{d\to\infty}\frac{1}{d}Q_p = \rho\Big(\frac{z_c^2}{b^2}\,m_\gamma'(z_c) + \frac{a}{b}\Big(z_c(z_c - 2z_a)\,m_\gamma(z_c) + z_a^2\,m_\gamma(z_a)\Big)\Big),$$

$$\operatorname{plim}_{d\to\infty}\frac{1}{d}\,\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \frac{\rho}{b}\Big(1 + z_c + z_c^2 m_\gamma(z_c)\Big).$$

Proof. First, we simplify Q_d:

$$\begin{aligned}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} &= \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{v}\\ &= \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-3/2}\boldsymbol{v}.\end{aligned}$$

Since $\mathbf{M}$ shares eigenvectors with $\boldsymbol{\Sigma}$, it commutes with $\boldsymbol{\Sigma}$, and we have that

$$\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-3/2} = \boldsymbol{\Sigma}^{-3}\mathbf{M}^{-1} = (a\boldsymbol{\Sigma}^3 + \boldsymbol{\Sigma}^2)^{-1}.$$

Hence, $Q_p = \boldsymbol{v}^\top(a\boldsymbol{\Sigma}^3 + \boldsymbol{\Sigma}^2)^{-1}\boldsymbol{v}$. Using Eqs. (28) and (30), we have

$$(a\boldsymbol{\Sigma}^3 + \boldsymbol{\Sigma}^2)^{-1}\boldsymbol{v} = \mathtt{vec}\left((a\boldsymbol{A}^3 + \boldsymbol{A}^2)^{-1}\boldsymbol{W}_d\right).$$

Using Eq. (33), we get

$$Q_p = \operatorname{tr}\left(\boldsymbol{W}_p(a\boldsymbol{A}^3 + \boldsymbol{A}^2)^{-1}\boldsymbol{W}_p\right) = \operatorname{tr}\left(\boldsymbol{W}_p^2(a\boldsymbol{A}^3 + \boldsymbol{A}^2)^{-1}\right).$$

Since $\boldsymbol{W}_p$ and $\boldsymbol{A}$ share eigenvectors, with eigenvalues λ_i and $c + b\lambda_i$, respectively,

$$Q_p = \sum_{i=1}^{p}\frac{\lambda_i^2}{(c+b\lambda_i)^2\,(a(c+b\lambda_i)+1)}.$$

As $p, M \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty}\frac{1}{d}Q_p = \frac{\rho}{ab^3}\int\frac{\lambda^2}{(\lambda + c/b)^2(\lambda + (ac+1)/ab)}\,\mathrm{d}\mu_\gamma(x).$$

Using Lemma 10 ($\mathcal{I}_3$ with $z_1 = z_c, z_2 = z_a$) gives the final expression for the first result.

For the second part, we have that

$$\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{v}.$$

Using Eqs. (28) and (30),

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = \boldsymbol{\Sigma}^{-1}\mathtt{vec}(\boldsymbol{W}_d) = \mathtt{vec}(\boldsymbol{A}^{-1}\boldsymbol{W}_d).$$

Using Eq. (33),

$$\boldsymbol{\theta}^\top\boldsymbol{\Sigma}\boldsymbol{\theta} = \mathtt{vec}(\boldsymbol{W}_p)^\top\mathtt{vec}(\boldsymbol{A}^{-1}\boldsymbol{W}_p) = \operatorname{tr}\left(\boldsymbol{W}_d\boldsymbol{A}^{-1}\boldsymbol{W}_d\right) = \operatorname{tr}\left(\boldsymbol{W}_d^2\boldsymbol{A}^{-1}\right).$$

Since $\boldsymbol{W}_p$ and $\boldsymbol{A}$ share eigenvectors, with eigenvalues λ_i and $c + b\lambda_i$, respectively,

$$\boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} = \sum_{i=1}^{p} \frac{\lambda_i^2}{c + b\lambda_i}.$$

As $p, M \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty} \frac{1}{d} \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} = \frac{\rho}{b} \int \frac{\lambda^2}{\lambda + c/b} \, \mathrm{d}\mu_\gamma(x).$$

Using Lemma 10 ($\mathcal{I}_1$ with $z_1 = 0, z_2 = z_c$), we get

$$\int \frac{\lambda}{\lambda + c/b} \, \mathrm{d}\mu_\gamma(x) = 1 + z_c m_\gamma(z_c).$$

Then using Lemma 10 ((v), with $\alpha_0 = 1, \alpha_1 = z_1 = z_c, \alpha_2 = 0$), we get the final expression. □

D.5 First Order Optimality Conditions of $D(\bar{u}, \bar{\beta})$

Theorem 3. *Consider $D(\bar{u}, \bar{\beta})$ as in Eq. (24). Let $a := \frac{\bar{\beta}\nu}{\bar{u}}$. Then, the solution of $\min_{\bar{u}} \max_{\bar{\beta}} D(\bar{u}, \bar{\beta})$ is given by*

$$\bar{u} = \frac{\bar{\beta}\nu}{a}, \quad a = \frac{\gamma\nu/\rho^2 - (c+1) + \sqrt{\left(\gamma\nu/\rho^2 - (c+1)\right)^2 + 4\gamma c\nu/\rho^2}}{2\gamma c(1 - \nu/\rho^2)}, \quad \gamma = \mu^{-1},$$

$$\bar{\beta}^2 = \frac{a^2\left(\sigma_n^2 + 1 - \rho + \rho c\right) + \rho(1 - a^2c^2) m_\gamma(z_a) - \rho a z_a^2 m_\gamma'(z_a)}{\rho^2 \left(1 - \frac{\nu}{\rho^2} - \frac{m_\gamma'(z_a)}{a^2}\right)}. \tag{34}$$

Proof. We derive the first order optimality conditions of $D(\bar{u}, \bar{\beta})$. First, we have

$$\begin{aligned} &\frac{\partial D(\bar{u}, \bar{\beta})}{\partial \bar{u}} = 0 \\ \implies &\frac{\bar{\beta}}{2} - \frac{\bar{\beta} c_\infty}{2\bar{u}^2} - \frac{\bar{\beta}^2}{2} \frac{\partial s_\infty(\bar{u}, \bar{\beta})}{\partial \bar{u}} - \frac{1}{2} \frac{\partial v_\infty(\bar{u}, \bar{\beta})}{\partial \bar{u}} = 0 \end{aligned} \tag{35}$$

Similarly, we have

$$\begin{aligned} &\frac{\partial D(\bar{u}, \bar{\beta})}{\partial \bar{\beta}} = 0 \\ \implies &\tfrac{\bar{u}}{2} + \frac{c_\infty}{2\bar{u}} - \bar{\beta} s_\infty(\bar{u}, \bar{\beta}) - \frac{\bar{\beta}^2}{2} \frac{\partial s_\infty(\bar{u}, \bar{\beta})}{\partial \bar{\beta}} - \frac{1}{2} \frac{\partial v_\infty(\bar{u}, \bar{\beta})}{\partial \bar{\beta}} = 0 \end{aligned} \tag{36}$$

We first simplify the limits v_∞, s_∞ before we get the partial derivatives. First recall that

$$\begin{aligned} \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}} &:= d^3 \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}} = \frac{1}{\tau} \left(\sigma_n^2 + \frac{1}{M} \sum_{i=1}^{M} \frac{\|\boldsymbol{w}_i\|^2}{d} \right) \mathbb{I}_{p^2} + \frac{1}{M} \sum_{i=1}^{M} (\boldsymbol{w}_i[\mathcal{S}] \boldsymbol{w}_i[\mathcal{S}]^\top) \otimes \mathbb{I}_p, \\ \overline{\boldsymbol{M}} &:= \tfrac{\boldsymbol{M}}{d^3} = \frac{\bar{\beta}\nu}{\bar{u}} \mathbb{I} + \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1} \\ \overline{\boldsymbol{\theta}}_\star &:= \frac{\boldsymbol{\theta}_\star}{d} = \frac{1}{dn} \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{H}_{\mathcal{S}_p} \boldsymbol{y}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1} \left(\frac{1}{M} \sum_{i=1}^{M} \boldsymbol{w}_i[\mathcal{S}] \otimes \boldsymbol{w}_i[\mathcal{S}] \right). \end{aligned}$$

Then, since $\lim_{d\to\infty} \frac{\|\boldsymbol{w}_i\|^2}{d} = 1$, using Lemma 2 with $\mathbf{M} = \overline{\boldsymbol{M}}$, $a = \frac{\bar{\beta}\nu}{\bar{u}}$, $b = 1$, $c = \frac{\sigma_n^2 + 1}{\tau}$, $\gamma = \mu^{-1}$, $z_a = -\frac{1}{a} - c$, we get

$$s_\infty(\bar{u}, \bar{\beta}) = \lim_{d\to\infty} d \operatorname{tr}(\boldsymbol{M}^{-1}) = \lim_{d\to\infty} \tfrac{1}{d^2} \operatorname{tr}(\overline{\boldsymbol{M}}^{-1}) = \rho^2 \left(\tfrac{1}{a} - \tfrac{1}{a^2} m_\gamma(z_a) \right). \tag{37}$$

Using Lemma 3 with $\mathbf{\Sigma} = \overline{\mathbf{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}$, $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}_\star$, we get

$$\begin{aligned} v_\infty(\bar{u},\bar{\beta}) &= \lim_{d\to\infty} \frac{1}{d}(\mathbf{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star)^\top \boldsymbol{M}^{-1}(\mathbf{\Sigma}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star) = \lim_{d\to\infty} \frac{1}{d}(\overline{\mathbf{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star)^\top \overline{\boldsymbol{M}}^{-1}(\overline{\mathbf{\Sigma}}^{-1/2}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star) \\ &= \rho(c^2 m'_\gamma(z_c) + ac(c+2z_a)m_\gamma(z_c) + az_a^2 m_\gamma(z_a)). \end{aligned}$$

As both $s_\infty(\bar{u},\bar{\beta})$ and $v_\infty(\bar{u},\bar{\beta})$ depend on $\bar{u},\bar{\beta}$ through a, we use the derivatives with respect to a to find the partial derivatives wrt $\bar{u},\bar{\beta}$.

$$\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial \bar{u}} = \frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\frac{\partial a}{\partial \bar{u}} = \frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\left(-\frac{\bar{\beta}\nu}{\bar{u}^2}\right).$$

$$\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial \bar{\beta}} = \frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\frac{\partial a}{\partial \bar{\beta}} = \frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\left(\frac{\nu}{\bar{u}}\right).$$

$$\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial \bar{u}} = \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\frac{\partial a}{\partial \bar{u}} = \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\left(-\frac{\bar{\beta}\nu}{\bar{u}^2}\right).$$

$$\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial \bar{\beta}} = \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\frac{\partial a}{\partial \bar{\beta}} = \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\left(\frac{\nu}{\bar{u}}\right).$$

Substituting back in Eq. (35)

$$\begin{aligned} \frac{\partial D(\bar{u},\bar{\beta})}{\partial \bar{u}} &= \frac{\bar{\beta}}{2} - \frac{\bar{\beta}c_\infty}{2\bar{u}^2} - \frac{\bar{\beta}^2}{2}\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial \bar{u}} - \frac{1}{2}\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial \bar{u}} \\ &= \frac{\bar{\beta}}{2} - \frac{\bar{\beta}c_\infty}{2\bar{u}^2} - \frac{\bar{\beta}^2}{2}\left(\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\right)\left(-\frac{\bar{\beta}\nu}{\bar{u}^2}\right) - \frac{1}{2}\left(\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\right)\left(-\frac{\bar{\beta}\nu}{\bar{u}^2}\right) \\ &= \frac{\bar{\beta}}{2} - \frac{\bar{\beta}c_\infty}{2\bar{u}^2} + \frac{\bar{\beta}^3\nu}{2\bar{u}^2}\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} + \frac{\bar{\beta}\nu}{2\bar{u}^2}\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a} = 0. \end{aligned}$$

Simplifying it further we get this is

$$\begin{aligned} & \bar{u}^2 - c_\infty + \bar{\beta}^2\nu\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} + \nu\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a} = 0 \\ \iff \quad & \bar{u}^2 = c_\infty - \nu\left(\bar{\beta}^2\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} + \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\right). \end{aligned} \tag{38}$$

Similarly, substituting in Eq. (36)

$$\begin{aligned} \frac{\partial D(\bar{u},\bar{\beta})}{\partial \bar{\beta}} &= \frac{\bar{u}}{2} + \frac{c_\infty}{2\bar{u}} - \bar{\beta}\, s_\infty(\bar{u},\bar{\beta}) - \frac{\bar{\beta}^2}{2}\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial \bar{\beta}} - \frac{1}{2}\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial \bar{\beta}} \\ &= \frac{\bar{u}}{2} + \frac{c_\infty}{2\bar{u}} - \bar{\beta}\, s_\infty(\bar{u},\bar{\beta}) - \frac{\bar{\beta}^2}{2}\left(\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a}\right)\left(\frac{\nu}{\bar{u}}\right) - \frac{1}{2}\left(\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\right)\left(\frac{\nu}{\bar{u}}\right) = 0. \end{aligned}$$

Simplifying it further we have

$$\bar{u} + \frac{c_\infty}{\bar{u}} - 2\bar{\beta}\, s_\infty(\bar{u},\bar{\beta}) - \frac{\nu}{\bar{u}}\left(\bar{\beta}^2\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} + \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\right) = 0.$$

Using Eq. (38), we get

$$\begin{aligned} 0 &= \bar{u} + \frac{c_\infty}{\bar{u}} - 2\bar{\beta}\, s_\infty(\bar{u},\bar{\beta}) - \frac{1}{\bar{u}}\left(c_\infty - \bar{u}^2\right) \\ &= 2\bar{u} - 2\bar{\beta}\, s_\infty(\bar{u},\bar{\beta}) \\ \iff \quad & \bar{u} = \bar{\beta}\, s_\infty(\bar{u},\bar{\beta}). \end{aligned} \tag{39}$$

Using Eq. (37) and simplifying Eq. (39), we get

$$\begin{aligned} & \bar{u} = \tfrac{\bar{\beta}\rho^2}{a}\left(1 - \tfrac{1}{a}m_\gamma\left(z_a\right)\right) \\ \iff \quad & \tfrac{\rho^2}{\nu}\left(1 - \tfrac{1}{a}m_\gamma\left(z_a\right)\right) = 1, \\ \iff \quad & \left(1 - \tfrac{\nu}{\rho^2}\right)a = m_\gamma\left(z_a\right). \end{aligned} \tag{40}$$

Eq. (40) can be simplified as follows. MP quadratic for $m_\gamma(z)$:

$$z\gamma m_\gamma(z)^2 + (z+\gamma-1)m_\gamma(z) + 1 = 0.$$

Substituting $z = z_a = -1/a - c, m_\gamma(z_a) = a(1-\nu)$, we have

$$\begin{aligned}
&\left(-\frac{1}{a} - c\right)\gamma a^2(1-\nu/\rho^2)^2 + \left(-\frac{1}{a} - c + \gamma - 1\right)a(1-\nu/\rho^2) + 1 = 0 \\
\iff & -(1+ac)\gamma a(1-\nu/\rho^2)^2 - (1+ac)(1-\nu/\rho^2) + (\gamma-1)a(1-\nu/\rho^2) + 1 = 0 \\
\iff & -\gamma c(1-\nu/\rho^2)^2 a^2 + (1-\nu/\rho^2)\big(\gamma\nu/\rho^2 - (c+1)\big)a + \nu/\rho^2 = 0 \\
\iff & c\gamma(1-\nu/\rho^2)a^2 + \big((c+1) - \gamma\nu/\rho^2\big)a - \tfrac{\nu/\rho^2}{1-\nu/\rho^2} = 0.
\end{aligned}$$

Solving for a using the quadratic formula finishes the proof for the first part.

Define

$$q_\infty(\bar{u}, \bar{\beta}) \;:=\; -\left(\bar{\beta}^2 \frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} + \frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a}\right).$$

Next to get a simplified form of this, we use $\frac{dz}{da} = \frac{1}{a^2}$ and get the partials

$$\begin{aligned}
s_\infty(\bar{u},\bar{\beta}) &= \rho^2\left(\frac{1}{a} - \frac{1}{a^2}m_\gamma(z_a)\right), \\
\frac{\partial s_\infty(\bar{u},\bar{\beta})}{\partial a} &= \rho^2\left(-\frac{1}{a^2} + \frac{2}{a^3}m_\gamma(z_a) - \frac{1}{a^4}m'_\gamma(z_a)\right).
\end{aligned}$$

$$\begin{aligned}
v_\infty(\bar{u},\bar{\beta}) &= \rho(c^2 m'_\gamma(-c) + ac(c+2z_a)m_\gamma(-c) + a(z_a)^2 m_\gamma(z_a)) \\
\frac{\partial v_\infty(\bar{u},\bar{\beta})}{\partial a} &= \rho(cm_\gamma(-c)\left(c + 2z_a + \frac{2}{a}\right) + z_a m_\gamma(z_a)\left(z_a + \frac{2}{a}\right) + \frac{z_a^2}{a}m'_\gamma(z_a)) \\
&= \rho(-c^2 m_\gamma(-c) + z_a m_\gamma(z_a)\left(\frac{1}{a} - c\right) + \frac{z_a^2}{a}m'_\gamma(z_a)).
\end{aligned}$$

$$\begin{aligned}
q_\infty(\bar{u},\bar{\beta}) &= -\Bigg[\bar{\beta}^2\rho^2\left(-\frac{1}{a^2} + \frac{2}{a^3}m_\gamma(z_a) - \frac{1}{a^4}m'_\gamma(z_a)\right) \\
&\qquad + \rho\left(-c^2 m_\gamma(-c) + z_a m_\gamma(z_a)\left(\frac{1}{a} - c\right) + \frac{z_a^2}{a}m'_\gamma(z_a)\right)\Bigg] \\
&= \frac{\bar{\beta}^2\rho^2}{a^2}\left(1 - \frac{2}{a}m_\gamma(z_a) + \frac{1}{a^2}m'_\gamma(z_a)\right) + \rho\left(c^2 m_\gamma(z_c) + \left(\frac{1}{a^2} - c^2\right)m_\gamma(z_a) - \frac{z_a^2}{a}m'_\gamma(z_a)\right).
\end{aligned} \tag{41}$$

Next, we simplify c_∞. Using Eq. (10) and Lemma 3 with $c = \frac{\sigma_n^2+1}{\tau}, b = 1$, we have

$$\begin{aligned}
c_\infty &= \lim_{d\to\infty}\frac{1}{d^2}\left(\sum_{i=1}^n d_i^2\right) = \lim_{d\to\infty}\frac{1}{d^2}\left(\sum_{i=1}^n \sigma_x^2\|\boldsymbol{w}_i\|^2 + n\sigma_n^2 - n\boldsymbol{\theta}_\star^\top \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\theta}_\star\right) \\
&= \lim_{d\to\infty}\frac{1}{d^2}\left(\sum_{i=1}^M \frac{n}{M}\sigma_x^2\|\boldsymbol{w}_i\|^2\right) - \nu\lim_{d\to\infty}\boldsymbol{\theta}_\star^\top\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}\boldsymbol{\theta}_\star + \nu\sigma_n^2 \\
&= \nu\lim_{d\to\infty}\frac{1}{Md}\left(\sum_{i=1}^M\|\boldsymbol{w}_i\|^2\right) - \nu\lim_{d\to\infty}\frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}\overline{\boldsymbol{\theta}}_\star + \nu\sigma_n^2 \\
&= \nu(1+\sigma_n^2) - \nu\rho(1 - c + c^2 m_\gamma(-c)) = \nu(\sigma_n^2 + 1 - \rho + \rho c - \rho c^2 m_\gamma(-c)).
\end{aligned} \tag{42}$$

We can write Eq. (38) as

$$\bar{u}^2 = c_\infty + \nu\, q_\infty(\bar{u}, \bar{\beta}),$$

Substituting c_∞ and q_∞ using Eqs. (41) and (42), and using $a = \frac{\bar{\beta}\nu}{\bar{u}}$, we get

$$\bar{\beta}_*^2 \nu = a^2(\sigma_n^2 + 1 - \rho + \rho c) + \bar{\beta}^2\rho^2 \left(1 - \frac{2}{a} m_\gamma(z_a) + \frac{1}{a^2} m'_\gamma(z_a)\right) + \rho\left(1 - a^2 c^2\right) m_\gamma(z_a) - \rho a z_a^2 m'_\gamma(z_a).$$

Solving for $\bar{\beta}^2$ gives Eq. (34).

□

E Proof for LS AO

The PO is

$$\min_{\boldsymbol{a}} \frac{1}{2}\|\boldsymbol{G}\boldsymbol{a} - \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}'\|^2 = \min_{\boldsymbol{a}} \max_{\boldsymbol{u}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{a} - \boldsymbol{u}^\top \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}' - \frac{1}{2}\|\boldsymbol{u}\|^2 .$$

By direct application of the CGMT, the AO is

$$\min_{\boldsymbol{a}} \max_{\boldsymbol{u}} \|\boldsymbol{u}\|\boldsymbol{g}^\top \boldsymbol{a} + \|\boldsymbol{a}\|\mathbf{h}^\top \boldsymbol{u} - \boldsymbol{u}^\top \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}' - \frac{1}{2}\|\boldsymbol{u}\|^2.$$

Let $\alpha = \|\boldsymbol{a}\|, \beta = \|\boldsymbol{u}\|$, and $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{u}}$ be the corresponding unit norm directions. Then the above can be written as

$$\min_{\alpha\geq 0, \hat{\boldsymbol{a}}} \max_{\beta\geq 0, \hat{\boldsymbol{u}}} \beta\alpha\boldsymbol{g}^\top \hat{\boldsymbol{a}} + \beta\hat{\boldsymbol{u}}^\top(\alpha\mathbf{h} - \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}') - \frac{1}{2}\beta^2.$$

Solving for $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{u}}$, the optimal is at

$$\hat{\boldsymbol{a}} = -\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}, \quad \hat{\boldsymbol{u}} = \frac{\alpha\mathbf{h} - \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}'}{\|\alpha\mathbf{h} - \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}'\|}.$$

Plugging these back, we have

$$\begin{aligned}
&\min_{\alpha\geq 0} \max_{\beta\geq 0} \beta\|\alpha\mathbf{h} - \boldsymbol{D}_{\boldsymbol{w}}\boldsymbol{g}'\| - \beta\alpha\|\boldsymbol{g}\| - \frac{1}{2}\beta^2 \\
&= \min_{\alpha\geq 0} \max_{\beta\geq 0} \beta\left(\sqrt{n\alpha^2 + \textstyle\sum_{i\in[n]} d_i^2} - \alpha\|\boldsymbol{g}\|\right) - \frac{1}{2}\beta^2 \\
&= \frac{1}{2}\left(\min_{\alpha\geq 0} \sqrt{n\alpha^2 + \sum_{i=1}^n d_i^2} - \alpha\|\boldsymbol{g}\|\right)_+^2
\end{aligned}$$

where $(\cdot)_+ = \max\{x, 0\}$. Thus, using Eq. (3),

$$\alpha_*^2 = \frac{\sum_{i=1}^n d_i^2}{n\left(\frac{n}{\|\boldsymbol{g}\|^2} - 1\right)} = \frac{\sum_{i=1}^n d_i^2}{\nu d^2\left(\frac{\nu p^2}{\rho^2\|\boldsymbol{g}\|^2} - 1\right)}.$$

Then, we get $\boldsymbol{a}_* = -\alpha_*\hat{\boldsymbol{g}}$. Using Eq. (23),

$$\kappa_\infty := \lim_{d\to\infty} \alpha_* = \sqrt{\frac{c_\infty}{\nu\left(\frac{\nu}{\rho^2} - 1\right)}}.$$

In the limit, $\boldsymbol{a}_* = \kappa_\infty \frac{\boldsymbol{g}}{\|\boldsymbol{g}\|}$.

F Asymptotic Risk

F.1 Asymptotic Risk of Minimum-Norm LGP

Lemma 4 (Term-1, OOD Loss)**.** *Using $\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{w}), \boldsymbol{u}, \boldsymbol{\theta}$ of the same form as Appendix D.2, let*

$$T_1(\boldsymbol{w}) := \frac{1}{d^2}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u},$$
$$T_2(\boldsymbol{w}) := \frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}, \qquad T_3(\boldsymbol{w}) := \frac{1}{d^{3/2}}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}.$$

Then, we have convergence to the following deterministic limits in probability

$$\begin{aligned}
&\mathrm{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_1(\boldsymbol{w})] = (1+b'+c')J_1,\\
&\mathrm{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_2(\boldsymbol{w})] = \frac{(1+b'+c')\rho}{b^2}\left(1+2z_c m_\gamma(z_c)+z_c^2 m'_\gamma(z_c)\right), \qquad (43)\\
&\mathrm{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_3(\boldsymbol{w})] = -(1+b'+c')\rho\left(-\frac{c(ac+2)}{b^3}m_\gamma(z_c)+\frac{a}{b}z_a^2 m_\gamma(z_a)+\frac{1}{b^2}z_c^2 m'_\gamma(z_c)\right), \qquad (44)
\end{aligned}$$

where

$$J_1 = \frac{\rho}{b^2}\left[\frac{\beta^2\rho}{a^2}\Big(bm_\gamma(z_a)-\frac{1}{a}m'_\gamma(z_a)\Big) - 2\,abz_c z_a\big(m_\gamma(z_c)-m_\gamma(z_a)\big)+z_c^2 m'_\gamma(z_c)+z_a^2 m'_\gamma(z_a)\right]. \qquad (45)$$

Proof. We first work with $T_1(\boldsymbol{w})$. Using Lemma 11, since $\boldsymbol{w}$ is independent of $\boldsymbol{g}, \boldsymbol{W}_p$, we get

$$\mathbb{E}_{\boldsymbol{w}}[T_1(\boldsymbol{w})] = \frac{c'+b'+1}{d^2}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-1}\boldsymbol{u}.$$

Since $\mathbf{M}$ and $\boldsymbol{\Sigma}$ share eigenvectors, they commute, and we have that $\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-1} = \boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}$. Further, we have that

$$\begin{aligned}
\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{u} &= \beta^2\boldsymbol{g}^\top(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2})\boldsymbol{g} + d\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}\\
&\quad - \beta d^{1/2}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{g} - \beta d^{1/2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}\\
&= \beta^2\boldsymbol{g}^\top(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2})\boldsymbol{g} + d\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-2}\boldsymbol{\theta} - 2\beta d^{1/2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-2}\boldsymbol{\theta}.
\end{aligned}$$

$$\begin{aligned}
J_1 &= \lim_{d\to\infty}\frac{1}{d^2}\boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{u}\\
&= \lim_{p\to\infty}\frac{\rho^2}{p^2}\beta^2\boldsymbol{g}^\top(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2})\boldsymbol{g} + \lim_{p\to\infty}\frac{\rho}{p}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-2}\boldsymbol{\theta} - \lim_{p\to\infty}\frac{\rho^{3/2}}{p^{3/2}}2\beta\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-2}\boldsymbol{\theta}\\
&= \lim_{p\to\infty}\beta^2\mathcal{I}'_1(p) + \lim_{p\to\infty}\mathcal{I}'_2(p), \qquad (46)
\end{aligned}$$

where

$$\mathcal{I}'_1(p) := \frac{1}{d^2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}), \quad \mathcal{I}'_2(p) := \lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-2}\boldsymbol{\theta},$$

For the first term as $\boldsymbol{g}\in\mathbb{R}^{p^2}$ we apply Lemma 14 with $d_1 = p^2$ and get

$$\lim_{p\to\infty}\frac{1}{p^2}\boldsymbol{g}^\top(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2})\boldsymbol{g} = \lim_{p\to\infty}\frac{1}{p^2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}).$$

Also, for the third term, from Lemma 14 it follows that

$$\lim_{p\to\infty}\frac{1}{p^{3/2}}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-2}\boldsymbol{\theta} = 0, \qquad (47)$$

as $\|\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-2}\boldsymbol{\theta}\| = O(\|\boldsymbol{\theta}\|) = O(\mathtt{vec}(\boldsymbol{W}_p)) = O(\sqrt{p})$.

For $\mathcal{I}_1'(p)$, using Eqs. (28), (29) and (31) we have that

$$\begin{aligned}\mathcal{I}_1'(p) &= \frac{\rho^2}{p^2}\operatorname{tr}((\boldsymbol{A}^{-1}\boldsymbol{B}^{-2})\otimes\mathbb{I}_p)\\ &= \frac{\rho^2}{p}\operatorname{tr}((c\mathbb{I}_p+b\boldsymbol{W}_p)^{-1}(a\mathbb{I}_p+(c\mathbb{I}_p+b\boldsymbol{W}_p)^{-1})^{-2})\\ &= \frac{\rho^2}{p}\sum_{i=1}^{p}\frac{1}{c+b\lambda_i}\frac{1}{\left(a+\frac{1}{c+b\lambda_i}\right)^2} = \frac{\rho^2}{p}\sum_{i=1}^{p}\frac{c+b\lambda_i}{\left(1+a(c+b\lambda_i)\right)^2}.\end{aligned}$$

As $p, M\to\infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{p\to\infty}\mathcal{I}_1'(p) = \frac{\rho^2}{a^2b}\int\frac{\lambda+c/b}{(\lambda+(1+ac)/ab)^2}\,d\mu_{\mathrm{MP},\gamma}(\lambda).$$

Using Lemma 10 ($\mathcal{I}_2$ with $z_1=z_c, z_2=z_a$) gives

$$\lim_{p\to\infty}\mathcal{I}_1'(p) = \frac{\rho^2}{a^2b}(m_\gamma(z_a)-\frac{1}{ab}m_\gamma'(z_a)). \tag{48}$$

For $\mathcal{I}_2'(p)$, since $\mathbf{M}$ and $\boldsymbol{\Sigma}$ share eigenvectors, and using Eqs. (28), (30), (31) and (33), we have

$$\begin{aligned}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-2}\boldsymbol{\theta} &= \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-2}\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{\Sigma}^{-4}\mathbf{M}^{-2}\boldsymbol{v}\\ &= \boldsymbol{v}^\top(\boldsymbol{A}^{-4}\boldsymbol{B}^{-2}\otimes\mathbb{I})\mathtt{vec}(\boldsymbol{W}_p) = \mathtt{vec}(\boldsymbol{W}_p)^\top\mathtt{vec}(\boldsymbol{A}^{-4}\boldsymbol{B}^{-2}\boldsymbol{W}_p)\\ &= \operatorname{tr}(\boldsymbol{W}_p^2\boldsymbol{A}^{-4}\boldsymbol{B}^{-2}).\end{aligned} \tag{49}$$

Using this, we get

$$\mathcal{I}_2'(p) = \frac{\rho}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(c+b\lambda_i)^4}\frac{1}{(a+(c+b\lambda_i)^{-1})^2} = \frac{\rho}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(c+b\lambda_i)^2(a(c+b\lambda_i)+1)^2}.$$

As $p, M\to\infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{p\to\infty}\mathcal{I}_2'(p) = \frac{\rho}{a^2b^4}\int\frac{\lambda^2}{(\lambda+c/b)^2(\lambda+(1+ac)/ab)^2}\mathrm{d}\mu_\gamma(\lambda)$$

Using Lemma 10 ($\mathcal{I}_7$ with $z_1=z_a, z_2=z_c$), we get

$$\lim_{d\to\infty}\mathcal{I}_2'(p) = \frac{\rho}{a^2b^4}\left[\frac{2z_az_c}{(z_a-z_c)^3}\big(m_\gamma(z_c)-m_\gamma(z_a)\big)+\frac{z_c^2}{(z_a-z_c)^2}m_\gamma'(z_c)+\frac{z_a^2}{(z_a-z_c)^2}m_\gamma'(z_a)\right]. \tag{50}$$

Using $z_a-z_c=-1/(ab)$ to simplify Eq. (50), and substituting it with Eq. (48) in Eq. (46) gives the final expression in Eq. (45).

Next, working with $T_2(\boldsymbol{w})$, using Lemma 11, we have

$$\mathbb{E}_{\boldsymbol{w}}\left[\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}\right] = \frac{(1+b'+c')}{d}\boldsymbol{\theta}^\top\boldsymbol{\theta} = \frac{(1+b'+c')}{d}\boldsymbol{v}^\top\boldsymbol{\Sigma}^{-2}\boldsymbol{v}.$$

Using Eqs. (28), (30) and (33), we have

$$\boldsymbol{v}^\top\boldsymbol{\Sigma}^{-2}\boldsymbol{v} = \mathtt{vec}(\boldsymbol{W}_p)^\top\mathtt{vec}(\boldsymbol{A}^{-2}\boldsymbol{W}_p) = \operatorname{tr}(\boldsymbol{W}_p^2(c\mathbb{I}_p+b\boldsymbol{W}_p)^{-2}). \tag{51}$$

Therefore,

$$\mathbb{E}_{\boldsymbol{w}}\left[\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}\right] = \frac{(1+b'+c')\rho}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(c+b\lambda_i)^2}.$$

As $p, M \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w}}\left[\frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}\right] = \frac{(1+b'+c')\rho}{b^2}\int \frac{\lambda^2}{(\lambda+c/b)^2}\mathrm{d}\mu_\gamma(\lambda). \tag{52}$$

Using Lemma 10 ($\mathcal{I}_6$ with $z_1 = z_c$) then gives the expression in Eq. (43).

Next, working with $T_3(\boldsymbol{w})$, using Lemma 11, we have that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}}[T_3(\boldsymbol{w})] &= \frac{1+b'+c'}{d^{3/2}}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} \\
&= \frac{1+b'+c'}{d^{3/2}}(\beta \boldsymbol{g} - d^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta})^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} \\
&= \frac{1+b'+c'}{d}\left(\beta d^{-1/2}\boldsymbol{g}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} - \boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}\right).
\end{aligned}$$

Using Lemma 14, we have that

$$\lim_{d\to\infty} \frac{1}{d^{3/2}}\boldsymbol{g}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} = \lim_{p\to\infty} \frac{\rho^{3/2}}{p^{3/2}}\boldsymbol{g}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta} = 0,$$

as $\|\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}\| = O(\|\boldsymbol{\theta}\|) = O(\|\boldsymbol{\Sigma}^{-1}\mathtt{vec}(\boldsymbol{W}_p)\| = O(\sqrt{p})$.

Using this we have

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w}}[T_3(\boldsymbol{w})] = -(1+b'+c')\lim_{d\to\infty} \mathcal{I}_3'(p), \quad \mathcal{I}_3'(p) = \frac{1}{d}\boldsymbol{\theta}^\top \boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\theta}. \tag{53}$$

Since $\boldsymbol{\Sigma}$ and $\mathbf{M}$ share eigenvectors, and using Eqs. (28), (30), (31) and (33), we have

$$\begin{aligned}
\mathcal{I}_3'(p) &= \frac{1}{d}\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1}\boldsymbol{v} = \frac{1}{d}\boldsymbol{v}^\top \boldsymbol{\Sigma}^{-3}\mathbf{M}^{-1}\boldsymbol{v} \\
&= \frac{1}{d}\mathtt{vec}(\boldsymbol{W}_p)^\top \mathtt{vec}(\boldsymbol{A}^{-3}\boldsymbol{B}^{-1}\boldsymbol{W}_p) = \frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^2\boldsymbol{A}^{-3}\boldsymbol{B}^{-1}) \\
&= \frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^2(c\mathbb{I}+b\boldsymbol{W}_p)^{-3}(a\mathbb{I}+(c\mathbb{I}+b\boldsymbol{W}_p)^{-1})^{-1}) \\
&= \frac{\rho}{p}\sum_{i=1}^{p}\frac{\lambda_i^2}{(c+b\lambda_i)^2(a(c+b\lambda_i)+1)}.
\end{aligned} \tag{54}$$

As $p, M \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding,

$$\lim_{d\to\infty} \mathcal{I}_3'(p) = \frac{\rho}{ab^2}\int \frac{\lambda^2}{(\lambda+c/b)^2\big(\lambda+(1+ac)/ab\big)}\, d\mu_\gamma(\lambda). \tag{55}$$

Using Lemma 10 ($\mathcal{I}_3$ with $z_1 = z_c, z_2 = z_a$) and simplifying gives the final expression in Eq. (44). □

Lemma 5 (Term-2, OOD Loss)**.** *Using $\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{u}, \boldsymbol{\theta}$ of the same form as Appendix D.2, consider the scalar random variable*

$$U_p(\boldsymbol{w}) := \frac{1}{d^{3/2}}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{w}[\mathcal{S}]\otimes\boldsymbol{w}[\mathcal{S}]) + \frac{1}{d}\boldsymbol{\theta}^\top(\boldsymbol{w}[\mathcal{S}]\otimes\boldsymbol{w}[\mathcal{S}]). \tag{56}$$

Then, it converges in probability as,

$$\mathrm{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[U_p(\boldsymbol{w})] = \frac{\rho}{b}\Big(1 + z_a m_\gamma(z_a)\Big).$$

Proof. Using $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w}[\mathcal{S}]\otimes\boldsymbol{w}[\mathcal{S}]] = \mathtt{vec}(\mathbb{I}_p)$ and since $\boldsymbol{w}$ is independent of $\boldsymbol{g}$ and $\boldsymbol{W}_p$, we have that

$$\mathbb{E}_{\boldsymbol{w}}[U_p(\boldsymbol{w})] = \frac{1}{d^{3/2}}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p) + \frac{1}{d}\boldsymbol{\theta}^\top \mathtt{vec}(\mathbb{I}_p).$$

Considering the first term, we have

$$\frac{1}{d^{3/2}}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p)=\frac{\beta}{d^{3/2}}\boldsymbol{g}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p)-\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p).$$

Using Lemma 14, we have that

$$\lim_{d\to\infty}\frac{\beta}{d^{3/2}}\boldsymbol{g}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p)=\lim_{p\to\infty}\frac{\beta\rho^{3/2}}{p^{3/2}}\boldsymbol{g}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\mathbb{I}_p)=0,$$

as $\|\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1}\mathtt{vec}(\mathbb{I}_p)\|=O(\sqrt{p})$. Since $\mathbf{M}$ and $\boldsymbol{\Sigma}$ share eigenvectors, they commute and we have that $\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}=\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-1}$. Therefore

$$\lim_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[U_p(\boldsymbol{w})]=-\lim_{d\to\infty}\mathcal{I}_1'(p)+\lim_{d\to\infty}\mathcal{I}_2'(p) \tag{57}$$

$$\mathcal{I}_1'(p)=\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-1}\mathtt{vec}(\mathbb{I}_p),\quad \mathcal{I}_2'(p)=\frac{1}{d}\boldsymbol{\theta}^\top\mathtt{vec}(\mathbb{I}_p). \tag{58}$$

For $\mathcal{I}_1'(p)$, using Eqs. (28), (30), (31) and (33) we have that

$$\begin{aligned}\mathcal{I}_1'(p)&=\frac{1}{d}\boldsymbol{v}^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-1}\mathtt{vec}(\mathbb{I}_p)=\frac{1}{d}\boldsymbol{v}^\top((\boldsymbol{A}^{-2}\boldsymbol{B}^{-1})\otimes\mathbb{I}_p)\mathtt{vec}(\mathbb{I}_p)=\frac{1}{d}\mathtt{vec}(\boldsymbol{W}_p)\mathtt{vec}(\boldsymbol{A}^{-2}\boldsymbol{B}^{-1})\\&=\frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p\boldsymbol{A}^{-2}\boldsymbol{B}^{-1})=\frac{\rho}{p}\sum_{i=1}^{p}\frac{\lambda_i}{(c+b\lambda_i)(a(c+b\lambda_i)+1)}\end{aligned}$$

As $p,M\to\infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty}\mathcal{I}_1'(p)=\frac{\rho}{ab^2}\int\frac{\lambda}{(\lambda+c/b)(\lambda+(1+ac)/ab)}\mathrm{d}\mu_\gamma(\lambda).$$

Using Lemma 10 ($\mathcal{I}_4$ with $z_1=z_c, z_2=z_a$), and $z_c-z_a=1/ab$ we get

$$\lim_{d\to\infty}\mathcal{I}_1'(p)=\frac{\rho}{b}(z_cm_\gamma(z_c)-z_am_\gamma(z_a)). \tag{59}$$

Next, for $\mathcal{I}_2'(p)$, using Eqs. (28), (30), (31) and (33), we have that

$$\mathcal{I}_2'(p)=\frac{1}{d}\boldsymbol{v}^\top\boldsymbol{\Sigma}^{-1}\mathtt{vec}(\mathbb{I}_p)=\frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p\boldsymbol{A}^{-1})=\frac{\rho}{p}\sum_{i=1}^{d}\frac{\lambda_i}{c+b\lambda_i}.$$

As $p,M\to\infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty}\mathcal{I}_2'(p)=\frac{\rho}{b}\int\frac{\lambda}{(\lambda+c/b)}\mathrm{d}\mu_\gamma(\lambda).$$

Using Lemma 10 ($\mathcal{I}_1$ with $z_1=z_c$), we get

$$\lim_{d\to\infty}\mathcal{I}_2'(p)=\frac{\rho}{b}(1+z_cm_\gamma(z_c)). \tag{60}$$

Substituting Eq. (59) and Eq. (60) in Eq. (57) finishes the proof. □

Lemma 6 (Term-1, ID Loss)**.** *Using $\boldsymbol{M},\boldsymbol{\Sigma},\boldsymbol{\Sigma}(\boldsymbol{w}_j),\boldsymbol{u},\boldsymbol{\theta}$ of the same form as Appendix D.2, let*

$$T_1(\boldsymbol{w}_j):=\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w}_j)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u},$$

$$T_2(\boldsymbol{w}_j):=\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}(\boldsymbol{w}_j)\boldsymbol{\theta},\qquad T_3(\boldsymbol{w}_j):=\frac{1}{d^{3/2}}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w}_j)\theta.$$

Then, we have convergence to the following deterministic limits in probability

$$\mathrm{plim}_{d\to\infty}\frac{1}{M}\sum_{j=1}^{M}T_1(w_j)=(b'+c')\,T_{1,a}+T_{1,b},$$

$$\mathrm{plim}_{d\to\infty}\frac{1}{M}\sum_{j=1}^{M}T_2(w_j)=(b'+c')T_{2,a}+T_{2,b}$$

$$\mathrm{plim}_{d\to\infty}\frac{1}{M}\sum_{j=1}^{M}T_3(w_j)=(b'+c')T_{3,a}+T_{3,b}$$

where $T_{1,a} = J_1$ from Eq. (45),

$$T_{1,b} = \frac{\rho}{b^2}\Bigg[-2abz_az_c(z_cm_\gamma(z_c) - z_am_\gamma(z_a)) + z_c^2(m_\gamma(z_c) + z_cm'_\gamma(z_c)) + z_a^2(m_\gamma(z_a) + z_am'_\gamma(z_a))\Bigg]$$
$$+ \frac{\rho^2\beta^2}{a^2b}\left(1 + \left(z_a - \frac{1}{ab}\right)m_\gamma(z_a) - \frac{1}{ab}z_am'_\gamma(z_a)\right),$$
$$T_{2,a} = \frac{\rho}{b^2}\left(1 + 2z_cm_\gamma(z_c) + z_c^2m'_\gamma(z_c)\right),$$
$$T_{2,b} = \frac{\rho}{b^2}\left(1 + 2z_c + 3z_c^2m_\gamma(z_c) + z_c^3m'_\gamma(z_c)\right),$$
$$T_{3,a} = \rho\left(-\frac{c(ac+2)}{b^3}m_\gamma(z_c) + \frac{a}{b}z_a^2m_\gamma(z_a) + \frac{1}{b^2}z_c^2m'_\gamma(z_c)\right),$$
$$T_{3,b} = \rho\left(-\frac{c(ac+2)}{b^3}(1 + z_cm_\gamma(z_c)) + \frac{a}{b}z_a^2(1 + z_am_\gamma(z_a)) + \frac{1}{b^2}z_c^2(m_\gamma(z_c) + z_cm'_\gamma(z_c))\right).$$

Proof. Recall

$$\boldsymbol{\Sigma}(\boldsymbol{w}_j) = \left(c'\frac{\|\boldsymbol{w}_j\|^2}{d} + b'\right)\mathbb{I}_{p^2} + (\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top) \otimes \mathbb{I}_p.$$

We have

$$T_1(w_j) = \frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w}_j)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}$$
$$= \underbrace{\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\left(c'\frac{\|\boldsymbol{w}_j\|^2}{d} + b'\right)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}}_{T_{1,a}(\boldsymbol{w}_j)} + \underbrace{\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}((\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top) \otimes \mathbb{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}}_{T_{1,b}(\boldsymbol{w}_j)}.$$

Let us first consider the limit of $T_{1,a}(\boldsymbol{w}_j)$. We have

$$\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{1,a}(\boldsymbol{w}_j) = \lim_{d\to\infty}\frac{1}{d^2M}\sum_{i=1}^M \boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\left(c'\frac{\|\boldsymbol{w}_j\|^2}{d} + b'\right)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}$$
$$= (b' + c')\lim_{d\to\infty}\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-1}\boldsymbol{u}, \tag{61}$$

where we use the fact that $\lim_{d\to\infty}\|\boldsymbol{w}\|^2/d = 1$ for $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbb{I}_d)$. The above limit is same as Eq. (45) derived in OOD loss.

Next, we solve the limit of $T_{1,b}(\boldsymbol{w}_j)$. We have

$$\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{1,b}(\boldsymbol{w}_j) = \lim_{d\to\infty}\frac{1}{d^2M}\sum_{i=1}^M \boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}((\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top) \otimes \mathbb{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}$$
$$= \lim_{d\to\infty}\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\left(\frac{1}{M}\sum_{j=1}^M(\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top) \otimes \mathbb{I}_p\right)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}$$
$$= \lim_{d\to\infty}\frac{1}{d^2}\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{W}_p \otimes \mathbb{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u}.$$

Using the definitions of $\boldsymbol{\Sigma}, \mathbf{M}$ we have

$$\boldsymbol{u}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{W}_p \otimes \mathbb{I}_p)\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{u} = \boldsymbol{u}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-2}\boldsymbol{W}_p \otimes \mathbb{I}_p\boldsymbol{u},$$

where we use the fact that $\boldsymbol{\Sigma}, \boldsymbol{M}, (\boldsymbol{W}_p \otimes \mathbb{I}_p)$ share Eigenvectors and hence commute. Plugging the definition of $\boldsymbol{u}$, we have

$$\boldsymbol{u}^\top\boldsymbol{M}^{-2}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_p \otimes \mathbb{I}_p\boldsymbol{u} = \beta^2\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{-2}\boldsymbol{W}_p \otimes \mathbb{I}_p\boldsymbol{g} + d\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\boldsymbol{M}^{-2}\boldsymbol{W}_p \otimes \mathbb{I}_p\boldsymbol{\theta}$$
$$- 2\beta d^{1/2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\boldsymbol{M}^{-2}\boldsymbol{W}_p \otimes \mathbb{I}_p\boldsymbol{\theta}.$$

Following similar steps to Eq. (47),

$$\lim_{d\to\infty}\frac{1}{d^{3/2}}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\boldsymbol{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p\boldsymbol{\theta}=\lim_{p\to\infty}\frac{\rho^{3/2}}{p^{3/2}}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-3/2}\mathbf{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p\boldsymbol{\theta}=0,$$

Using Lemma 14 again, we get

$$\frac{1}{d^2}\lim_{d\to\infty}\boldsymbol{g}^\top(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p)\boldsymbol{g}=\lim_{p\to\infty}\frac{\rho^2}{p^2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p).$$

Following similar steps to the calculation of Eq. (48) in OOD loss, we get the following integral (with an additional λ) factor in the limit $p, M\to\infty$

$$\lim_{p\to\infty}\frac{\rho^2}{p^2}\operatorname{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p)=\frac{\rho^2}{a^2b}\int\lambda\frac{\lambda+c/b}{(\lambda+(1+ac)/ab)^2}\,d\mu_{\mathrm{MP},\gamma}(\lambda).$$

Simplifying this using result in Eq. (48) and Lemma 10 (v) we get

$$\frac{\rho^2}{a^2b}\int\lambda\frac{\lambda+c/b}{(\lambda+(1+ac)/ab)^2}\,d\mu_{\mathrm{MP},\gamma}(\lambda)=\frac{\rho^2}{a^2b}\left(1+\left(z_a-\frac{1}{ab}\right)m_\gamma(z_a)-\frac{1}{ab}z_am'_\gamma(z_a)\right).$$

Again, using similar steps to result derived in Eq. (50), we get

$$\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-2}\boldsymbol{M}^{-2}\boldsymbol{W}_p\otimes\mathbb{I}_p\boldsymbol{\theta}=\frac{\rho}{a^2b^4}\int\lambda\frac{\lambda^2}{(\lambda+c/b)^2(\lambda+(1+ac)/ab)^2}\mathrm{d}\mu_{\mathrm{MP},\gamma}(\lambda)$$

Using Lemma 10 (v) with Eq. (50) gives the following expression for the above integral

$$\frac{\rho}{a^2b^4}\Bigg[\frac{2z_az_c}{(z_a-z_c)^3}(z_cm_\gamma(z_c)-z_am_\gamma(z_a))+\frac{z_c^2}{(z_a-z_c)^2}(m_\gamma(z_c)+z_cm'_\gamma(z_c))$$
$$+\frac{z_a^2}{(z_a-z_c)^2}(m_\gamma(z_a)+z_am'_\gamma(z_a))\Bigg] \tag{62}$$

Combining Eq. (61) and Eq. (62), gives the result. Next, we consider $T_2(\boldsymbol{w}_j)$. It is easy to see that it decomposes to the following similar to $T_1(\boldsymbol{w}_j)$

$$\begin{aligned}T_2(\boldsymbol{w}_j)&=\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}(\boldsymbol{w}_j)\boldsymbol{\theta}\\&=\underbrace{\frac{1}{d}\boldsymbol{\theta}\left(c'\frac{\|\boldsymbol{w}_j\|^2}{d}+b'\right)\boldsymbol{\theta}}_{T_{2,a}(\boldsymbol{w}_j)}+\underbrace{\frac{1}{d}\boldsymbol{\theta}^\top((\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top)\otimes\mathbb{I}_p)\boldsymbol{\theta}}_{T_{2,b}(\boldsymbol{w}_j)}.\end{aligned}$$

The limiting average $\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_2(\boldsymbol{w}_j)$, similar to the average for $T_1(\boldsymbol{w}_j)$, it is easy to see that $\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{2,a}(\boldsymbol{w}_j)$ uses the same integral as derived in Eq. (43) for the OOD loss.

The term $T_{2,b}(\boldsymbol{w}_j)$ average in the limit

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{dM}\sum_{j=1}^M T_{2,b}(\boldsymbol{w}_j)&=\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top\left(\frac{1}{M}\sum_{j=1}^M(\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top)\otimes\mathbb{I}_p\right)\boldsymbol{\theta}\\&=\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top(\boldsymbol{W}_p\otimes\mathbb{I}_p)\boldsymbol{\theta}.\end{aligned}$$

Following similar steps to the calculation in Eq. (51), we get the following trace expression (with an additional $\boldsymbol{W}_p$ factor)

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top(\boldsymbol{W}_p\otimes\mathbb{I}_p)\boldsymbol{\theta}&=\lim_{d\to\infty}\frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^3(c\mathbb{I}_p+b\boldsymbol{W}_p)^{-2})\\&=\rho\int\lambda\frac{\lambda^2}{(\lambda+c/b)^2}\mathrm{d}\mu_\gamma(\lambda),\\&=\frac{\rho}{b^2}\left(1+2z_cm_\gamma(z_c)+3z_c^2m_\gamma(z_c)+z_c^3m'_\gamma(z_c)\right).\end{aligned}$$

where the integral follows from calculation in Eq. (52). The final equality follows by combining Lemma 10 (v) with integral solved in Eq. (52).

Next, for $T_3(\boldsymbol{w}_j)$, we get a similar decomposition to $T_1(\boldsymbol{w}_j)$ and $T_2(\boldsymbol{w}_j)$

$$T_3(\boldsymbol{w}_j) = \underbrace{\boldsymbol{u}^\top \mathbf{M}^{-1}\Big(c'\frac{\|\boldsymbol{w}_j\|^2}{d} + b'\Big)\boldsymbol{\theta}}_{T_{3,a}(\boldsymbol{w}_j)} + \underbrace{\boldsymbol{u}^\top \mathbf{M}^{-1}((\boldsymbol{w}_j[\mathcal{S}]\boldsymbol{w}_j[\mathcal{S}]^\top)\otimes \mathbb{I}_p)\boldsymbol{\theta}}_{T_{3,b}(\boldsymbol{w}_j)}.$$

The limiting average $\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_3(\boldsymbol{w}_j)$, similar to the limiting averages computed for $T_1(\boldsymbol{w}_j)$ and $T_2(\boldsymbol{w}_j)$, it is easy to see that $\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{3,a}(\boldsymbol{w}_j)$ uses the integral Eq. (55).

For $\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{3,b}(\boldsymbol{w}_j)$, following similar calculations to Eq. (54), we have the expression below (with an additional $\boldsymbol{W}_p$ factor)

$$\begin{aligned}
\lim_{d\to\infty}\frac{1}{d^2M}\sum_{j=1}^M T_{3,b}(\boldsymbol{w}_j) &= \frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^3\boldsymbol{A}^{-3}\boldsymbol{B}^{-1}) \\
&= \frac{\rho}{ab^2}\int \lambda \frac{\lambda^2}{(\lambda+c/b)^2(\lambda+(1+ac)/ab)}\,d\mu_\gamma(\lambda) \\
&= \rho\left(-\frac{c(ac+2)}{b^3}(1+z_c m_\gamma(z_c)) + \frac{a}{b}z_a^2(1+z_a m_\gamma(z_a)) + \frac{1}{b^2}z_c^2(m_\gamma(z_c)+z_c m_\gamma'(z_c))\right),
\end{aligned}$$

where the second equality follows similar steps from Eq. (54) to Eq. (55), and the last inequality follows by combining the result in Eq. (55) with Lemma 10 (v). □

Lemma 7 (Term-2, ID Loss)**.** *Using $\boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{w}), \boldsymbol{u}, \boldsymbol{\theta}$ of the same form as Appendix D.2, let*

$$T_2(\boldsymbol{w}_j) := \frac{1}{d^{3/2}}\boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{w}_j[\mathcal{S}]\otimes\boldsymbol{w}_j[\mathcal{S}]) + \frac{1}{d}\boldsymbol{\theta}^\top(\boldsymbol{w}_j[\mathcal{S}]\otimes\boldsymbol{w}_j[\mathcal{S}]).$$

Then, we have convergence to the following deterministic limit in probability

$$\operatorname{plim}_{d\to\infty}\frac{1}{M}\sum_{j=1}^M T_2(\boldsymbol{w}_j) = -\frac{\rho}{b}(z_c(1+z_c m_\gamma(z_c)) - z_a(1+z_a m_\gamma(z_a))) + \frac{\rho}{b}(1+z_c(1+z_c m_\gamma(z_c)))$$

Proof. Consider the average of the first term. Using $\frac{1}{M}\sum_{j=1}^M \boldsymbol{w}_j[\mathcal{S}]\otimes\boldsymbol{w}_j[\mathcal{S}] = \mathtt{vec}(\boldsymbol{W}_p)$, we have

$$\frac{1}{Md^{3/2}}\sum_{j=1}^M \boldsymbol{u}^\top \mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{w}_j[\mathcal{S}]\otimes\boldsymbol{w}_j[\mathcal{S}]) = \frac{\beta}{d^{3/2}}\boldsymbol{g}^\top\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\boldsymbol{W}_p) - \frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\boldsymbol{W}_p).$$

Term 1 above is 0 in the limit $d\to\infty$, using Lemma 14. Term 2 simplifies to

$$\begin{aligned}
\frac{1}{d}\boldsymbol{\theta}^\top\boldsymbol{\Sigma}^{-1/2}\mathbf{M}^{-1}\boldsymbol{\Sigma}^{-1/2}\mathtt{vec}(\boldsymbol{W}_p) &= \frac{1}{d}\mathtt{vec}(\boldsymbol{W}_p)^\top\boldsymbol{\Sigma}^{-2}\mathbf{M}^{-1}\mathtt{vec}(\boldsymbol{W}_p) = \frac{1}{d}\mathtt{vec}(\boldsymbol{W}_p)^\top\boldsymbol{A}^{-2}\boldsymbol{B}^{-1}\otimes\mathbb{I}_p\mathtt{vec}(\boldsymbol{W}_p) \\
&= \frac{1}{d}\mathtt{vec}(\boldsymbol{W}_p)^\top\mathtt{vec}(\boldsymbol{A}^{-2}\boldsymbol{B}^{-1}\boldsymbol{W}_p) \\
&= \frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^2\boldsymbol{A}^{-2}\boldsymbol{B}^{-1}).
\end{aligned}$$

Here, second equality follows from the definition of $\boldsymbol{\theta}$ and the fact that $\mathbf{M}$ and $\boldsymbol{\Sigma}$ commute. Third and fourth equality use the definition of $\boldsymbol{\Sigma}, \mathbf{M}$ and Eq. (30) and Eq. (31). In the limit

$$\begin{aligned}
\lim_{d\to\infty}\frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^2\boldsymbol{A}^{-2}\boldsymbol{B}^{-1}) &= \frac{\rho}{ab^2}\int\frac{\lambda}{(\lambda+c/b)(\lambda+(1+ac)/ab)}\mathrm{d}\mu_\gamma(\lambda) \\
&= \frac{\rho}{b}(z_c(1+z_c m_\gamma(z_c)) - z_a(1+z_a m_\gamma(z_a))),
\end{aligned}$$

where second equality follows by Lemma 10 with Eq. (59). Next, for term 2

$$\begin{aligned}\lim_{d\to\infty}\frac{1}{d}\boldsymbol{\theta}^\top \mathtt{vec}(\boldsymbol{W}_p) &= \lim_{d\to\infty}\frac{1}{d}\operatorname{tr}(\boldsymbol{W}_p^2\boldsymbol{A}^{-1})\\ &= \frac{\rho}{b}\int \lambda\frac{\lambda}{(\lambda+c/b)}\mathrm{d}\mu_\gamma(\lambda)\\ &= \frac{\rho}{b}(1+z_c(1+z_c m_\gamma(z_c))).\end{aligned}$$

The first equality follows by using definition of $\boldsymbol{\theta}, \boldsymbol{\Sigma}$ and Eq. (30) and Eq. (31). Second equality follows by the fact the empirical distribution of $\boldsymbol{W}_p$ converges to Marchenko-Pastur law μ_γ in the limit, and the final equality uses Lemma 10 with Eq. (60). □

We restate Theorem 1 below followed by the proof.

Theorem 4 (Asymptotic Risk of Minimum-Norm LGP)**.** *Consider the minimum norm estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ for the linear Gaussian equivalent problem in Definition 2 under the overparameterized regime ($\nu < \rho^2$). Under the asymptotic scaling defined in Eq. (3), as $d \to \infty$, the ID and OOD risks converge in probability to the following deterministic limits:*

$$\operatorname{plim}_{d\to\infty}\mathcal{L}_{OOD}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = \left(1-\tfrac{\nu}{\rho^2}\right)\rho(1+ac)(1-c(1+a+ac)) + (1+a+ac)(\sigma_n^2+\rho c+1-\rho), \tag{63}$$

$$\operatorname{plim}_{d\to\infty}\mathcal{L}_{ID}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = \frac{\nu}{a^2}\bar{\beta}^2. \tag{64}$$

Proof. From Eq. (17) and using Eqs. (18) to (21), we have

$$\boldsymbol{a} = \boldsymbol{M}^{-1}(\beta\boldsymbol{g} - \boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\theta}_\star) = \frac{1}{d}\overline{\boldsymbol{M}}^{-1}(\bar{\beta}\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star).$$

Substituting this in Lemma 1, we get Term-1(a) as

$$\frac{1}{d^2}(\bar{\beta}_*\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star)^\top\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{M}}^{-1}(\bar{\beta}_*\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star).$$

Similarly, Term-1(c) is

$$\boldsymbol{a}^\top\boldsymbol{\Sigma}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}_\star = \frac{1}{d^{3/2}}(\bar{\beta}_*\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star)^\top\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star$$

Similarly, Term-2 is

$$\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}^\top\mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X)y_q] = \frac{1}{d^{3/2}}(\bar{\beta}_*\boldsymbol{g} - d^{1/2}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\theta}}_\star)^\top\overline{\boldsymbol{M}}^{-1}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}(\boldsymbol{w}\otimes\boldsymbol{w}) + \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top\boldsymbol{w}\otimes\boldsymbol{w}.$$

OOD Loss. Using these, and Lemma 4 and Lemma 5 with $\boldsymbol{\Sigma} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}$, $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}_\star$, $\mathbf{M} = \overline{\boldsymbol{M}}$, $b = 1$, $b' + c' = c =$, $\gamma = \mu^{-1}$, $z_a = -\frac{1}{a} - c$, $z_c = -c$, and defining $m_a := m_\gamma(z_a)$, $m_a' := m_\gamma'(z_a)$, $m_c := m_\gamma(-c)$, $m_c' := m_\gamma'(-c)$, we have

$$\begin{aligned}\mathcal{L}_{\text{OOD}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = &\underbrace{(1+c)\left(\bar{\beta}^2\rho^2\left(\frac{1}{a^2}m_a - \frac{1}{a^3}m_a'\right) + \rho(c^2 m_c' + 2z_a ac(m_c - m_a) + z_a^2 m_a')\right)}_{\text{Term-1(a)}}\\ &+ \underbrace{(1+c)\rho\left(1 - 2cm_c + c^2 m_c'\right)}_{\text{Term-1(b)}} \underbrace{-2(1+c)\rho\left[-c(ac+2)m_c + az_a^2 m_a + c^2 m_c'\right]}_{\text{Term-1(c)}}\\ &\underbrace{-2\rho(1+z_a m_a)}_{\text{Term-2}} + 1 + \sigma_n^2.\end{aligned}$$

Using the expression for $\bar{\beta}$ from Eq. (34), Term-1(a) can be simplified as:

$$(1+c)\Big(a\big(\sigma_n^2+1-\rho+\rho c\big)+\rho z_a^2 a m_a+\rho c^2 m_c'-2\rho(1+ac)cm_c\Big)$$

Combining all terms, we get

$$\begin{aligned}\mathcal{L}_{\text{OOD}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) &= m_a(-z_a)\rho(-az_a(1+c)+2az_a(1+c)+2)+m_c(1+c)\rho(2z_a ac-2c+2c(ac+2))\\ &\quad+m_c'(1+c)\rho c^2(1+1-2)+(1+c)a(\sigma_n^2+1-\rho+\rho c)+\rho(1+c-2)+1+\sigma_n^2.\end{aligned}$$

Simplifying this and using $m_a = a(1-\nu/\rho^2)$, we get Eq. (63).

ID Loss. Using the expressions for Term-1(a), Term-1(c) and Term-2 computed above, Lemmas 1, 6 and 7, we have

$$\begin{aligned}\mathcal{L}_{\text{ID}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) &= \underbrace{c\Big(\bar{\beta}^2\rho^2\left(\frac{1}{a^2}m_a-\frac{1}{a^3}m_a'\right)+\rho(c^2m_c'+2z_a ac(m_c-m_a)+z_a^2 m_a')\Big)}_{\text{Term-1(a)-old part}}\\
&+\underbrace{\bar{\beta}^2\rho^2\left(\frac{1}{a^2}(1+z_a m_a)-\frac{1}{a^3}(m_a+z_a m_a')\right)+\rho(c^2(m_c+z_c m_c')+2z_a ac(z_c m_c-z_a m_a)+z_a^2(m_a+z_a m_a'))}_{\text{Term-1(a)-new part}}\\
&+\underbrace{c\rho\left(1-2cm_c+c^2\,m_c'\right)}_{\text{Term-1(b)-old}}+\underbrace{\rho(1-2c+3c^2m_c-c^3m_c')}_{\text{Term-1(b)-new}}\\
&-\underbrace{2c\rho\left[-c(ac+2)m_c+az_a^2m_a+c^2m_c'\right]}_{\text{Term-1(c)-old}}-\underbrace{2\rho\Big[a^{-1}+az_a^3m_a+(3c^2+ac^3)m_c-c^3m_c'\Big]}_{\text{Term-1(c)-new}}\\
&\underbrace{-2\rho(1-c+c^2m_c-a^{-1}-z_c^2m_c+z_a^2m_a)}_{\text{Term-2-new}}+1+\sigma_n^2.\end{aligned}$$

Combining terms, we have

$$\begin{aligned}\mathcal{L}_{\text{ID}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) &= \bar{\beta}^2\rho^2\frac{1}{a^2}\left((z_a+c)\left(m_a-\frac{1}{a}m_a'\right)+1-\frac{1}{a}m_a\right)+m_a\rho z_a\left(-2ac^2-2z_a ac+z_a-2acz_a-2az_a^2-2z_a\right)\\
&\quad+\rho m_a'\left(cz_a^2+z_a^3\right)+\rho m_c\left(2ac^2z_a+c^2-2z_a ac^2-2c^2+3c^2+2c^2(ac+2)-2(3c^2+ac^3)\right)\\
&\quad+\rho m_c'\left(c^3-c^3+c^3-c^3-2c^3+2c^3\right)+1+\sigma_n^2+\rho\left(-2(1-c)+1-2c+c\right)\\
&= \bar{\beta}^2\rho^2\frac{1}{a^2}\left(-\frac{1}{a}\left(m_a-\frac{1}{a}m_a'\right)+\frac{\nu}{\rho^2}\right)+\frac{\rho(1-a^2c^2)}{a^2}m_a-\frac{\rho z_a^2}{a}m_a'+\sigma_n^2+1-\rho+\rho c.\end{aligned}$$

Using the expression for $\bar{\beta}$ from Eq. (34), we simplify and get Eq. (64).

□

F.2 Asymptotic Risk of LS LGP

Lemma 8 (OOD Loss). *Using $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{w}), \boldsymbol{\theta}$ of the same form as Appendix D.2, let*

$$T_1(\boldsymbol{w}) := \frac{1}{\|\boldsymbol{g}\|^2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\Sigma}^{-1/2}\boldsymbol{g},$$

$$T_2(\boldsymbol{w}) := \frac{1}{\sqrt{d}\|\boldsymbol{g}\|}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}(\boldsymbol{w})\boldsymbol{\theta}, \qquad T_3(\boldsymbol{w}) := \frac{1}{\sqrt{d}\|\boldsymbol{g}\|}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{w}\otimes\boldsymbol{w}).$$

Then, we have convergence to the following deterministic limits in probability

$$\text{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_1(\boldsymbol{w})] = \frac{1+b'+c'}{b}m_\gamma(z_c), \qquad \text{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_2(\boldsymbol{w})] = 0, \qquad \text{plim}_{d\to\infty}\mathbb{E}_{\boldsymbol{w}}[T_3(\boldsymbol{w})] = 0.$$

Proof. We first work with $T_1(\boldsymbol{w})$. Using Lemma 11, since $\boldsymbol{w}$ is independent of $\boldsymbol{g}, \boldsymbol{W}_p$, we get

$$\mathbb{E}_{\boldsymbol{w}}[T_1(\boldsymbol{w})] = \frac{(c'+b'+1)}{\|\boldsymbol{g}\|^2}\boldsymbol{g}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{g}.$$

Using Lemma 14, and since $\lim_{d\to\infty} \frac{p}{\|\boldsymbol{g}\|} = 1$, we have that

$$\lim_{d\to\infty} \frac{1}{\|\boldsymbol{g}\|^2} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{g} = \lim_{d\to\infty} \frac{1}{p^2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1}).$$

Using Eqs. (28) and (29), we have

$$\frac{1}{p^2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1}) = \frac{1}{p} \operatorname{tr}(\boldsymbol{A}^{-1}) = \frac{1}{p} \sum_{i=1}^{p} \frac{1}{c + b\lambda_i}. \tag{65}$$

As $p \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ, yielding

$$\lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w}}[T_1(\boldsymbol{w})] = \frac{1}{b} \int \frac{1}{\lambda + c/b} \mathrm{d}\mu_\gamma(\lambda) = \frac{1}{b} m_\gamma(z_c).$$

Next, for $T_2(\boldsymbol{w})$, using Lemma 11, since $\boldsymbol{w}$ is independent of $\boldsymbol{g}, \boldsymbol{W}_p$, we have that

$$\mathbb{E}_{\boldsymbol{w}}[T_2(\boldsymbol{w})] = \frac{(c' + b' + 1)}{\sqrt{d}\|\boldsymbol{g}\|} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\theta}.$$

Using Lemma 14, it follows that

$$\lim_{d\to\infty} \frac{1}{p^{3/2}} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\theta} = 0,$$

which concludes the proof for the second part.

Finally, working with $T_3(\boldsymbol{w})$, Using $\mathbb{E}_{\boldsymbol{w}}[\boldsymbol{w} \otimes \boldsymbol{w}] = \mathtt{vec}(\mathbb{I}_p)$ and since $\boldsymbol{w}$ is independent of $\boldsymbol{g}$ and $\boldsymbol{W}_p$, we have that

$$\mathbb{E}_{\boldsymbol{w}}[T_3(\boldsymbol{w})] = \frac{1}{\sqrt{d}\|\boldsymbol{g}\|} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \mathtt{vec}(\mathbb{I}_p).$$

Using Lemma 14, it follows that

$$\lim_{d\to\infty} \frac{1}{p^{3/2}} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \mathtt{vec}(\mathbb{I}_p) = 0,$$

which concludes the proof for this part. □

Lemma 9 (ID Loss)**.** *Using $\boldsymbol{\Sigma}, \boldsymbol{\Sigma}(\boldsymbol{w}_j), \boldsymbol{u}, \boldsymbol{\theta}$ of the same form as Appendix D.2, let*

$$T_1(\boldsymbol{w}_j) := \frac{1}{\|\boldsymbol{g}\|^2} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}(\boldsymbol{w}_j) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{g},$$

$$T_2(\boldsymbol{w}_j) := \frac{1}{\sqrt{d}\|\boldsymbol{g}\|} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}(\boldsymbol{w}_j) \boldsymbol{\theta}, \qquad T_3(\boldsymbol{w}_j) := \frac{1}{\sqrt{d}\|\boldsymbol{g}\|} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{w}_j \otimes \boldsymbol{w}_j).$$

Then, we have convergence to the following deterministic limits in probability

$$\mathrm{plim}_{d\to\infty} \frac{1}{M} \sum_{j=1}^{M} T_1(\boldsymbol{w}_j) = \frac{1}{b}(1 + z_c m_\gamma(z_c)), \qquad \mathrm{plim}_{d\to\infty} \frac{1}{M} \sum_{j=1}^{M} T_2(\boldsymbol{w}_j) = 0, \qquad \mathrm{plim}_{d\to\infty} \frac{1}{M} \sum_{j=1}^{M} T_3(\boldsymbol{w}_j) = 0.$$

Proof. We first work with $T_1(\boldsymbol{w}_j)$. The average is

$$\frac{1}{M} \sum_{j=1}^{M} T_1(\boldsymbol{w}_j) = \frac{1}{M} \frac{1}{\|\boldsymbol{g}\|^2} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} \left(c' \tfrac{\|\boldsymbol{w}_j\|^2}{d} + b' \right) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{g} + \frac{1}{\|\boldsymbol{g}\|^2} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{W}_p \otimes \mathbb{I}_p) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{g}.$$

This just follows using definition of $\boldsymbol{\Sigma}(\boldsymbol{w}_j)$. Next, in the limit $d \to \infty$, term 1 follows using Eq. (65). The limit of second quantity above

$$\begin{aligned}
\lim_{d\to\infty} \frac{1}{\|\boldsymbol{g}\|^2} \boldsymbol{g}^\top \boldsymbol{\Sigma}^{-1/2} (\boldsymbol{W}_p \otimes \mathbb{I}_p) \boldsymbol{\Sigma}^{-1/2} \boldsymbol{g} &= \lim_{d\to\infty} \frac{1}{p^2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1/2} (\boldsymbol{W}_p \otimes \mathbb{I}_p) \boldsymbol{\Sigma}^{-1/2}) \\
&= \lim_{p\to\infty} \frac{1}{p^2} \operatorname{tr}(\boldsymbol{\Sigma}^{-1} \boldsymbol{W}_p \otimes \mathbb{I}_p)) \\
&= \lim_{p\to\infty} \frac{1}{p} \operatorname{tr}(\boldsymbol{A}^{-1} \boldsymbol{W}_p) \\
&= \frac{1}{b} \int \frac{\lambda}{\lambda + c/b} \mathrm{d}\mu_\gamma(\lambda) = \frac{1}{b}(1 + z_c m_\gamma(z_c)).
\end{aligned}$$

Here, we use commutation of $\boldsymbol{\Sigma}$ and $\boldsymbol{W}_p \otimes \mathbb{I}_p$ and properties Eq. (29), Eq. (31). The final equality follows as when $p, M \to \infty$, the empirical spectral distribution of $\boldsymbol{W}_p$ converges to the Marchenko-Pastur law μ_γ.

Next, the limiting averages of $T_2(\boldsymbol{w}_j)$ and $T_3(\boldsymbol{w}_j)$, using Lemma 14, and similar steps used in Lemma 8 both evaluate to 0. □

We restate Theorem 2 below followed by its proof.

Theorem 5 (Asymptotic Risk of LS LGP)**.** *Consider the least-squares estimator $\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}$ for the linear Gaussian equivalent problem in the underparameterized regime ($\nu > \rho^2$). Under the asymptotic scaling in Eq. (3), as $d \to \infty$, the ID and OOD risks converge in probability to:*

$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{OOD}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) = m_\gamma(z_c)(\kappa_\infty^2(1+c) - 2c^2\rho) + (1+c)\rho c^2 m'_\gamma(z_c) + \sigma_n^2 + 1 - \rho + \rho c,$$

$$\operatorname{plim}_{d\to\infty} \mathcal{L}_{ID} = \tfrac{\nu}{\rho^2}\kappa_\infty^2,$$

where $\kappa_\infty := \sqrt{\frac{\sigma_n^2 + 1 - \rho + \rho c - \rho c^2 m_c}{\frac{\nu}{\rho^2} - 1}}$.

Proof. Substituting $\boldsymbol{a} = \frac{\kappa_\infty}{\|\boldsymbol{g}\|}\boldsymbol{g}$ and using Eq. (26) and Eq. (27) in Eq. (25), we have that

$$\begin{aligned}\mathcal{L}(\boldsymbol{a};\boldsymbol{w}) = \frac{\kappa_\infty^2}{\|\boldsymbol{g}\|^2}\boldsymbol{g}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\boldsymbol{g} + \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top \overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star + 2\frac{\kappa_\infty}{\sqrt{d}\|\boldsymbol{g}\|}\boldsymbol{g}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}\overline{\boldsymbol{\Sigma}}(\boldsymbol{w})\overline{\boldsymbol{\theta}}_\star \\ - 2\left(\frac{\kappa_\infty}{d^{1/2}\|\boldsymbol{g}\|}\boldsymbol{g}^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{H}_{\mathcal{S}_p}}^{-1/2}(\boldsymbol{w}\otimes\boldsymbol{w}) + \frac{1}{d}\overline{\boldsymbol{\theta}}_\star^\top(\boldsymbol{w}\otimes\boldsymbol{w})\right) + 1 + \sigma_n^2.\end{aligned}$$

Using Lemma 8, Eq. (43), Eq. (60), and $b = 1, c' + b' = c$, we have that

$$\begin{aligned}\mathcal{L}_{\text{OOD}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) &:= \lim_{d\to\infty} \mathbb{E}_{\boldsymbol{w}}[\mathcal{L}(\boldsymbol{a};\boldsymbol{w})] \\ &= \kappa_\infty^2(1+c)m_\gamma(z_c) + (1+c)\rho\left(1 + 2z_c m_\gamma(z_c) + z_c^2 m'_\gamma(z_c)\right) - 2\rho(1 + z_c m_\gamma(z_c)) + 1 + \sigma_n^2 \\ &= m_\gamma(z_c)(\kappa_\infty^2(1+c) - 2c^2\rho) + (1+c)\rho c^2 m'_\gamma(z_c) + \sigma_n^2 + 1 - \rho + \rho c,\end{aligned}$$

where using Eq. (42),

$$\kappa_\infty^2 = \frac{c_\infty}{\nu\left(\frac{\nu}{\rho^2} - 1\right)} = \frac{\sigma_n^2 + 1 - \rho + \rho c - \rho c^2 m_\gamma(z_c)}{\frac{\nu}{\rho^2} - 1}.$$

Similarly, using Lemma 9, we have

$$\begin{aligned}\mathcal{L}_{\text{ID}}(\hat{\boldsymbol{\theta}}_{\mathcal{S}_p}) &:= \lim_{d\to\infty} \frac{1}{M}\sum_{i=1}^{M} \mathcal{L}(\boldsymbol{a};\boldsymbol{w}_i)] \\ &= \kappa_\infty^2\left(cm_\gamma(z_c) + 1 - cm_\gamma(z_c)\right) + c\rho\left(1 - 2cm_\gamma(z_c) + c^2 m'_\gamma(z_c)\right) \\ &+ \rho(1 - 2c + 3c^2 m_\gamma(z_c) - c^3 m'_\gamma(z_c)) - 2\rho(1 - c + c^2 m_\gamma(z_c)) + 1 + \sigma_n^2 \\ &= \kappa_\infty^2 - \rho c^2 m_\gamma(z_c) + \sigma_n^2 + 1 - \rho + \rho c = \tfrac{\nu}{\rho^2}\kappa_\infty^2.\end{aligned}$$

□

G Helper Lemmas

Lemma 10. *Let $z_1, z_2, z_3, \alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}$ be constants and $f(\cdot)$ be a function. Using*

$$m_\gamma(z) = \int \frac{1}{\lambda - z}\,\mathrm{d}\mu_\gamma(\lambda), \qquad m'_\gamma(z) = \int \frac{1}{(\lambda - z)^2}\,\mathrm{d}\mu_\gamma(\lambda),$$

(i) $\mathcal{I}_1 := \int \frac{\lambda - z_1}{\lambda - z_2} \, \mathrm{d}\mu_\gamma(\lambda) = 1 + (z_2 - z_1) m_\gamma(z_2)$,

(ii) $\mathcal{I}_2 := \int \frac{\lambda - z_1}{(\lambda - z_2)^2} \, \mathrm{d}\mu_\gamma(\lambda) = m_\gamma(z_2) + (z_2 - z_1) m'_\gamma(z_2)$,

(iii) $\mathcal{I}_3 := \int \frac{\lambda^2}{(\lambda - z_1)^2(\lambda - z_2)} \, \mathrm{d}\mu_\gamma(\lambda) = \frac{1}{(z_1 - z_2)^2}(z_1(z_1 - 2z_2) m_\gamma(z_1) + z_2^2 m_\gamma(z_2) + z_1^2(z_1 - z_2) m'_\gamma(z_1)$,

(iv) $\mathcal{I}_4 := \int \frac{\lambda}{(\lambda - z_1)(\lambda - z_2)} \, \mathrm{d}\mu_\gamma(\lambda) = \frac{1}{z_1 - z_2}(z_1 m_\gamma(z_1) - z_2 m_\gamma(z_2))$,

(v) *if* $\int f(\lambda) \, \mathrm{d}\mu_\gamma(\lambda) = \alpha_0 + \alpha_1 m_\gamma(z_1) + \alpha_2 m'_\gamma(z_1)$, *then*

$$\int \lambda f(\lambda) \, \mathrm{d}\mu_\gamma(\lambda) = \alpha_0 + \alpha_1 + (\alpha_1 z_1 + \alpha_2) m_\gamma(z_1) + \alpha_2 z_1 m'_\gamma(z_1),$$

(vi) $\mathcal{I}_6 := \int \frac{\lambda^2}{(\lambda - z_1)^2} \, \mathrm{d}\mu_\gamma(\lambda) = 1 + 2 z_1 m_\gamma(z_1) + z_1^2 m'_\gamma(z_1)$,

(vii) $\mathcal{I}_7 := \int \frac{\lambda^2}{(\lambda - z_1)^2(\lambda - z_2)^2} \, \mathrm{d}\mu_\gamma(\lambda) = \frac{1}{(z_1 - z_2)^2} \left(z_1^2 m'_\gamma(z_1) + z_2^2 m'_\gamma(z_2) - \frac{2 z_1 z_2}{z_1 - z_2} \left(m_\gamma(z_1) - m_\gamma(z_2) \right) \right)$.

Proof. The proof relies on obtaining partial fraction decompositions and using the definitions of $m_\gamma(z), m'_\gamma(z)$.

First, for $\mathcal{I}_1$, note that

$$\frac{\lambda - z_1}{\lambda - z_2} = 1 + \frac{z_2 - z_1}{\lambda - z_2}, \quad \mathcal{I}_1 = 1 + (z_2 - z_1) m_\gamma(z_2).$$

Next, for $\mathcal{I}_2$, we have

$$\frac{\lambda - z_1}{(\lambda - z_2)^2} = \frac{1}{\lambda - z_2} + \frac{z_2 - z_1}{(\lambda - z_2)^2} \implies \mathcal{I}_2 = m_\gamma(z_2) + (z_2 - z_1) m'_\gamma(z_2).$$

Next, for $\mathcal{I}_3$, we have

$$\frac{\lambda^2}{(\lambda - z_1)^2(\lambda - z_2)} = \frac{z_1^2}{z_1 - z_2} \frac{1}{(\lambda - z_1)^2} + \frac{z_1(z_1 - 2z_2)}{(z_1 - z_2)^2} \frac{1}{\lambda - z_1} + \frac{z_2^2}{(z_1 - z_2)^2} \frac{1}{\lambda - z_2},$$

which gives $\mathcal{I}_3 = \frac{1}{(z_1 - z_2)^2}(z_1(z_1 - 2z_2) m_\gamma(z_1) + z_2^2 m_\gamma(z_2) + z_1^2(z_1 - z_2) m'_\gamma(z_1)$.

Next, for $\mathcal{I}_4$, we have

$$\frac{\lambda}{(\lambda - z_1)(\lambda - z_2)} = \frac{z_1}{z_1 - z_2} \frac{1}{\lambda - z_1} + \frac{z_2}{z_2 - z_1} \frac{1}{\lambda - z_2},$$

which gives $\mathcal{I}_4 = \frac{z_1}{z_1 - z_2} m_\gamma(z_1) + \frac{z_2}{z_2 - z_1} m_\gamma(z_2)$.

For the next part, we have

$$\alpha_1 m_\gamma(z_1) + \alpha_2 m'_\gamma(z_1) = \alpha_0 + \alpha_1 \int \frac{1}{\lambda - z_1} \, \mathrm{d}\mu_\gamma(\lambda) + \alpha_2 \int \frac{1}{(\lambda - z_1)^2} \, \mathrm{d}\mu_\gamma(\lambda)$$

This gives

$$\begin{aligned} \int \lambda f(\lambda) \, \mathrm{d}\mu_\gamma(\lambda) &= \alpha_0 \int \lambda \, \mathrm{d}\mu_\gamma(\lambda) + \alpha_1 \int \frac{\lambda}{\lambda - z_1} \, \mathrm{d}\mu_\gamma(\lambda) + \alpha_2 \int \frac{\lambda}{(\lambda - z_1)^2} \, \mathrm{d}\mu_\gamma(\lambda) \\ &= \alpha_0 + \alpha_1(1 + z_1 m_\gamma(z_1)) + \alpha_2(m_\gamma(z_1) + z_1 m_\gamma(z_1)), \end{aligned}$$

where we used $\mathcal{I}_1$ for the first term and $\mathcal{I}_2$ for second term. Simplifying this finishes the proof for this part.

Next, for $\mathcal{I}_6$, we have

$$\frac{\lambda^2}{(\lambda - z_1)^2} = \left(1 + \frac{z_1}{\lambda - z_1}\right)^2 = 1 + \frac{2 z_1}{\lambda - z_1} + \frac{z_1^2}{(\lambda - z_1)^2},$$

which gives $\mathcal{I}_6 = 1 + 2z_1 m_\gamma(z_1) + z_1^2 m'_\gamma(z_1)$.

Next, for $\mathcal{I}_7$, we have

$$\begin{aligned}\frac{\lambda^2}{(\lambda - z_1)^2(\lambda - z_2)^2} &= \frac{1}{(z_1 - z_2)^2}\left(\frac{z_1}{\lambda - z_1} - \frac{z_2}{\lambda - z_2}\right)^2 \\ &= \frac{1}{(z_1 - z_2)^2}\left(\frac{z_1^2}{(\lambda - z_1)^2} + \frac{z_2^2}{(\lambda - z_2)^2} - \frac{2z_1 z_2}{z_1 - z_2}\left(\frac{1}{\lambda - z_1} - \frac{1}{\lambda - z_2}\right)\right),\end{aligned}$$

which gives $\mathcal{I}_7 = \frac{1}{(z_1 - z_2)^2}\Big(z_1^2 m'_\gamma(z_1) + z_2^2 m'_\gamma(z_2) - \frac{2z_1 z_2}{z_1 - z_2}\left(m_\gamma(z_1) - m_\gamma(z_2)\right)\Big)$. □

Lemma 11. $\mathbb{E}_{\boldsymbol{w}}[\Sigma(\boldsymbol{w})] = (c' + b' + 1)\mathbb{I}_{d^2}$.

Proof. Since $\boldsymbol{w} \sim \mathcal{N}(0, I_d)$, we have $\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top] = \mathbb{I}_d$ and $\mathbb{E}[\|\boldsymbol{w}\|^2/d] = 1$, hence

$$\mathbb{E}_{\boldsymbol{w}}[\Sigma(\boldsymbol{w})] = \Big(c'\mathbb{E}[\tfrac{\|\boldsymbol{w}\|^2}{d}] + b'\Big)\mathbb{I}_{d^2} + \mathbb{E}[(\boldsymbol{w}\boldsymbol{w}^\top) \otimes \mathbb{I}_d] = (c' + b' + 1)\mathbb{I}_{d^2}.$$

□

Lemma 12 (Gaussian tail bound)**.** *If $X \sim \mathcal{N}(0, \sigma^2)$, then for all $t > 0$,*

$$\Pr(|X| \geq t) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

Lemma 13 (Sub-exponential tail bound)**.** *If X_i are i.i.d., zero mean, sub-exponential random variables, then there exist universal constants $K, c_0 > 0$ such that $\|X_i\|_{\psi_1} \leq K$ and*

$$\Pr\Big(\Big|\sum_{i=1}^d a_i X_i\Big| \geq t\Big) \leq 2\exp\Big(-c_0 \min\Big\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\Big\}\Big) \quad \textit{for all } t > 0.$$

Lemma 14. *Let $\boldsymbol{a}_1 \in \mathbb{R}^{d_1}$ and $\boldsymbol{A}_1 \in \mathbb{R}^{d_1 \times d_1}$, s.t. $\|\boldsymbol{a}_1\| = O(1/d_1^q)$, $q > 0$ and $\lambda_{\max}(\boldsymbol{A}_1) = O(1)$. Then,*

$$\lim_{d_1 \to \infty} \boldsymbol{g}^\top \boldsymbol{a}_1 = 0$$
$$\lim_{d_1 \to \infty} \frac{1}{d_1}\boldsymbol{g}^\top \boldsymbol{A}\boldsymbol{g} = \lim_{d_1 \to \infty} \frac{1}{d_1}\operatorname{tr}(\boldsymbol{A}_1).$$

Proof. For the first, we know that if $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_{d_1})$, then $\boldsymbol{g}^\top \boldsymbol{a}_1 \sim \mathcal{N}(0, \|\boldsymbol{a}_1\|^2)$. Using standard Gaussian tail bounds from Lemma 12 with $\|\boldsymbol{a}_1\| = O(1/d_1^q)$, we have $\lim_{d_1 \to \infty} \boldsymbol{g}^\top \boldsymbol{a}_1 = 0$.

Next, let $\boldsymbol{A}_1 = \boldsymbol{V}\Lambda\boldsymbol{V}^\top$ be the Eigenvalue decomposition of $\boldsymbol{A}_1$. We have

$$\boldsymbol{g}^\top \boldsymbol{A}_1 \boldsymbol{g} = (\boldsymbol{V}^\top \boldsymbol{g})^\top \Lambda (\boldsymbol{V}^\top \boldsymbol{g}) = \tfrac{1}{d_1}\sum_{i=1}^d \lambda_i \tilde{g}_i{}^2,$$

where $\boldsymbol{V}^\top \boldsymbol{g} =: \tilde{\boldsymbol{g}} \sim \mathcal{N}(\boldsymbol{0}, \mathbb{I}_{d_1})$. We know that $X_i := \tilde{g}_i{}^2 - 1$ are 0 mean i.i.d sub-exponential random variables for all $i \in [d_1]$. Applying sub-exponential tail bound from Lemma 13 to $\sum_{i=1}^d \lambda_i X_i$, we have

$$\begin{aligned}\Pr\left(\left|\frac{1}{d_1}\sum_{i=1}^{d_1} \lambda_i X_i\right| \geq t\right) &= 2\exp\Big(-c_0 d_1 \min\Big\{\frac{d_1 t^2}{\|\Lambda\|_F^2}, \frac{t}{\lambda_{\max}}\Big\}\Big) \quad \text{for all } t > 0. \\ &\leq 2\exp\Big(-c_0 d_1 \min\Big\{\frac{t^2}{C^2}, \frac{t}{C}\Big\}\Big),\end{aligned}$$

where inequality uses $\lambda_i = \Theta(1)$. Therefore

$$\lim_{d_1 \to \infty} \frac{1}{d_1}\sum_{i=1}^{d_1} \lambda_i X_i = \lim_{d_1 \to \infty} \frac{1}{d_1}(\boldsymbol{g}^\top \boldsymbol{A}_1 \boldsymbol{g} - \operatorname{tr}(\boldsymbol{A}_1)) = 0,$$

where we use that $\sum_{i=1}^{d_1} \lambda_i = \operatorname{tr}(\Lambda) = \operatorname{tr}(\boldsymbol{A}_1)$. □

Lemma 15. *Let $\boldsymbol{x} \in \mathbb{R}^d$ have i.i.d. entries $x_i \sim \mathcal{N}(0, \sigma_x^2)$, and let $\boldsymbol{w} \in \mathbb{R}^d$ be fixed. For a subset $\mathcal{S} \subseteq [d]$ with $|\mathcal{S}| = p$, denote by $\boldsymbol{x}[\mathcal{S}] \in \mathbb{R}^p$ and $\boldsymbol{w}[\mathcal{S}] \in \mathbb{R}^p$ the corresponding coordinate sub-vectors. Then*

$$\mathbb{E}\big[(\boldsymbol{w}^\top \boldsymbol{x})^2\, \boldsymbol{x}[\mathcal{S}]\, \boldsymbol{x}[\mathcal{S}]^\top\big] = \sigma_x^4 \|\boldsymbol{w}\|^2\, \mathbb{I}_p \;+\; 2\sigma_x^4\, \boldsymbol{w}[\mathcal{S}] \boldsymbol{w}[\mathcal{S}]^\top.$$

Proof. Write $\boldsymbol{x} = \sigma_x \boldsymbol{g}$ where $\boldsymbol{g} \sim \mathcal{N}(0, \mathbb{I}_d)$ has independent standard normal entries. Then

$$(\boldsymbol{w}^\top \boldsymbol{x})^2\, x_i x_j = \sigma_x^4 \big(\boldsymbol{w}^\top \boldsymbol{g}\big)^2 g_i g_j = \sigma_x^4 \Big(\sum_{k=1}^d w_k g_k\Big)^2 g_i g_j.$$

Expanding the square and taking expectations,

$$\mathbb{E}\big[(\boldsymbol{w}^\top \boldsymbol{x})^2\, x_i x_j\big] = \sigma_x^4 \sum_{k,\ell=1}^d w_k w_\ell\, \mathbb{E}[g_k g_\ell g_i g_j].$$

By Isserlis' theorem for a zero-mean Gaussian vector,

$$\mathbb{E}[g_k g_\ell g_i g_j] = \delta_{k\ell}\delta_{ij} + \delta_{ki}\delta_{\ell j} + \delta_{kj}\delta_{\ell i},$$

where δ_{ab} is the Kronecker delta. Hence

$$\begin{aligned}
\mathbb{E}\big[(\boldsymbol{w}^\top \boldsymbol{x})^2\, x_i x_j\big] &= \sigma_x^4 \sum_{k,\ell=1}^d w_k w_\ell \big(\delta_{k\ell}\delta_{ij} + \delta_{ki}\delta_{\ell j} + \delta_{kj}\delta_{\ell i}\big) \\
&= \sigma_x^4 \Big(\delta_{ij} \sum_{k=1}^d w_k^2 + w_i w_j + w_j w_i\Big) \\
&= \sigma_x^4 \Big(\delta_{ij} \|\boldsymbol{w}\|^2 + 2 w_i w_j\Big).
\end{aligned}$$

Restricting to indices $i, j \in \mathcal{S}$ and collecting these entries into a $p \times p$ matrix yields

$$\mathbb{E}\big[(\boldsymbol{w}^\top \boldsymbol{x})^2\, \boldsymbol{x}[\mathcal{S}] \boldsymbol{x}[\mathcal{S}]^\top\big] = \sigma_x^4 \|\boldsymbol{w}\|^2 \mathbb{I}_p + 2\sigma_x^4 \boldsymbol{w}[\mathcal{S}] \boldsymbol{w}[\mathcal{S}]^\top,$$

as claimed. □

Lemma 16. *Under the setup described in Section 2, where $\mathcal{S} \subseteq [d]$ be any subset of indices of size p, $\boldsymbol{w} \in \mathbb{R}^d$ be a task vector, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T, \boldsymbol{x}_q \in \mathbb{R}^d$ be i.i.d. samples from $\mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I}_d)$, labels $y_t = \boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \sigma_n^2)$, and*

$$\widehat{\boldsymbol{w}}_{\mathrm{avg}}[\mathcal{S}] := \frac{1}{T} \sum_{t=1}^T y_t\, \boldsymbol{x}_t[\mathcal{S}] \in \mathbb{R}^p, \quad \mathbf{h}_{\mathcal{S}_p}(X) := \mathrm{vec}\big(\boldsymbol{x}_q[\mathcal{S}]\, \widehat{\boldsymbol{w}}_{\mathrm{avg}}[\mathcal{S}]^\top\big).$$

Then the following hold:

$$\underbrace{\mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X) \mathbf{h}_{\mathcal{S}_p}(X)^\top]}_{\textit{Term-I}} = \frac{\sigma_x^4}{T} \left[(\sigma_x^2 \|\boldsymbol{w}\|^2 + \sigma_n^2) \mathbb{I}_{p^2} + (T+1)\sigma_x^2 (\boldsymbol{w}[\mathcal{S}] \boldsymbol{w}[\mathcal{S}]^\top) \otimes \mathbb{I}_p\right],$$

$$\underbrace{\mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q]}_{\textit{Term-II}} = \sigma_x^4 \boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}].$$

Proof. We first simplify Term-I. We have that

$$\mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X) \mathbf{h}_{\mathcal{S}_p}(X)^\top] = \mathbb{E}_{X,\epsilon}\left[(\mathtt{vec}(\boldsymbol{x}_q[\mathcal{S}] \hat{\boldsymbol{w}}_{\mathrm{avg}}[\mathcal{S}]^\top)(\mathtt{vec}(\boldsymbol{x}_q[\mathcal{S}] \hat{\boldsymbol{w}}_{\mathrm{avg}}[\mathcal{S}]^\top))^\top\right].$$

Using Eq. (32), and since $\boldsymbol{x}_q$ is independent of $\hat{\boldsymbol{w}}_{\mathrm{avg}} = \frac{1}{T}\sum_t y_t \boldsymbol{x}_t$, separating the expectation, we get

$$\mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X) \mathbf{h}_{\mathcal{S}_p}(X)^\top] = \frac{1}{T^2} \underbrace{\mathbb{E}\left[\left(\sum_{t=1}^T y_t \boldsymbol{x}_t[\mathcal{S}]\right)\left(\sum_{t'=1}^T y_{t'} \boldsymbol{x}_{t'}[\mathcal{S}]\right)^\top\right]}_{\text{Term-I(i)}} \otimes \underbrace{\mathbb{E}[\boldsymbol{x}_q[\mathcal{S}] \boldsymbol{x}_q[\mathcal{S}]^\top]}_{\text{Term-1(ii)}}$$

For Term-I(ii), since $\boldsymbol{x}_q[S] \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I}_p)$, we have $\mathbb{E}[\boldsymbol{x}_q[\mathcal{S}]\boldsymbol{x}_q[\mathcal{S}]^\top] = \sigma_x^2 \mathbb{I}_p$. Next, we simplify Term-I(i).

$$\mathbb{E}\left[\left(\sum_{t=1}^{T} y_t \boldsymbol{x}_t[\mathcal{S}]\right)\left(\sum_{t'=1}^{T} y_{t'} \boldsymbol{x}_{t'}[\mathcal{S}]\right)^\top\right] = \sum_{t=1}^{T}\sum_{t'=1}^{T} \mathbb{E}[y_t y_{t'} \boldsymbol{x}_t[\mathcal{S}]\boldsymbol{x}_{t'}[\mathcal{S}]^\top]$$
$$= \sum_{t=1}^{T} \mathbb{E}[y_t^2 \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_t[\mathcal{S}]^\top] + \sum_{t\neq t'} \mathbb{E}[y_t y_{t'} \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_{t'}[\mathcal{S}]^\top]$$

First, looking at $t \neq t'$ terms, we have:

$$\begin{aligned}\mathbb{E}[y_t y_{t'} \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_{t'}[\mathcal{S}]^\top] &= \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon_t)\ (\boldsymbol{w}^\top \boldsymbol{x}_{t'} + \epsilon_{t'})\, \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_{t'}[\mathcal{S}]^\top] \\ &= \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_t)\, \boldsymbol{x}_t[\mathcal{S}]]\ \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_{t'})\, \boldsymbol{x}_{t'}[\mathcal{S}]^\top] \\ &= \sigma_x^4 \Pi_{p\times d} \boldsymbol{w}\ \boldsymbol{w}^\top \Pi_{p\times d}^\top \\ &= \sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top,\end{aligned}$$

where the second equality follows by independence and $\mathbb{E}[\epsilon] = 0$, and for the third equality, we use the definition of $\Pi_{p\times d}$, and that $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I})$.

Next, consider the terms with $t = t'$:

$$\begin{aligned}\mathbb{E}[y_t^2 \boldsymbol{x}_t \boldsymbol{x}_t^\top] &= \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon_t)^2 \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_t[\mathcal{S}]^\top] \\ &= \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_t)^2\, \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_t[\mathcal{S}]^\top] + \mathbb{E}[\epsilon_t^2\, \boldsymbol{x}_t[\mathcal{S}]\, \boldsymbol{x}_t[\mathcal{S}]^\top] \\ &= \sigma_x^4 \|\boldsymbol{w}\|^2\, \mathbb{I}_p + 2\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top + \sigma_n^2 \sigma_x^2 \mathbb{I}_p \\ &= (\sigma_x^4 \|\boldsymbol{w}\|^2 + \sigma_x^2 \sigma_n^2)\mathbb{I}_p + 2\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top,\end{aligned}$$

where the third equality follows by using Lemma 15 and that $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbb{I}_d)$.

Combining the T diagonal terms and $T(T-1)$ off-diagonal terms, we have

$$\begin{aligned}\mathbb{E}\left[\left(\sum_{t=1}^{T} y_t \boldsymbol{x}_t[\mathcal{S}]\right)\left(\sum_{t'=1}^{T} y_{t'} \boldsymbol{x}_{t'}[\mathcal{S}]\right)^\top\right] &= T\left[(\sigma_x^4 \|\boldsymbol{w}\|^2 + \sigma_x^2 \sigma_n^2)\mathbb{I}_p + 2\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top\right] + T(T-1)\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top \\ &= T(\sigma_x^4 \|\boldsymbol{w}\|^2 + \sigma_x^2 \sigma_n^2)\mathbb{I}_p + (2T + T^2 - T)\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top \\ &= T(\sigma_x^4 \|\boldsymbol{w}\|^2 + \sigma_x^2 \sigma_n^2)\mathbb{I}_p + (T^2 + T)\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top.\end{aligned}$$

Combining this resultant Term-1(i) with Term-1(ii), we get Term-1:

$$\begin{aligned}\mathbb{E}_{X,\epsilon}[\mathbf{h}_{\mathcal{S}_p}(X)\mathbf{h}_{\mathcal{S}_p}(X)^\top] &= \frac{1}{T^2}\left[T(\sigma_x^4 \|\boldsymbol{w}\|^2 + \sigma_x^2 \sigma_n^2)\mathbb{I}_p + (T^2 + T)\sigma_x^4 \boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top\right] \otimes \sigma_x^2 \mathbb{I}_p \\ &= \frac{\sigma_x^4}{T}\left[(\sigma_x^2 \|\boldsymbol{w}\|^2 + \sigma_n^2)\mathbb{I}_{p^2} + (T+1)\sigma_x^2 (\boldsymbol{w}[\mathcal{S}]\boldsymbol{w}[\mathcal{S}]^\top) \otimes \mathbb{I}_p\right].\end{aligned}$$

Next, let us simplify Term-II:

$$\begin{aligned}\mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q] &= \mathbb{E}_{X,\epsilon}\left[\mathtt{vec}(\boldsymbol{x}_q[\mathcal{S}]\hat{\boldsymbol{w}}_{\text{avg}}[\mathcal{S}]^\top)(\boldsymbol{w}^\top \boldsymbol{x}_q + \epsilon)\right] \\ &= \frac{1}{T}\mathbb{E}\underbrace{\left[\sum_t y_t \boldsymbol{x}_t[\mathcal{S}]\right]}_{\text{Term-II(i)}} \otimes \mathbb{E}\underbrace{\left[\boldsymbol{x}_q[\mathcal{S}]\boldsymbol{x}_q^\top\right]\boldsymbol{w}}_{\text{Term-II(ii)}}.\end{aligned}$$

First, let us simplify Term-II(i). We have

$$\begin{aligned}\mathbb{E}\left[\sum_t y_t \boldsymbol{x}_t[\mathcal{S}]\right] &= \mathbb{E}\left[\sum_t (\boldsymbol{w}^\top \boldsymbol{x}_t + \epsilon)\boldsymbol{x}_t[\mathcal{S}]\right] \\ &= \sum_t \mathbb{E}\left[\boldsymbol{x}_t[\mathcal{S}]\boldsymbol{x}_t^\top\right]\boldsymbol{w} = T\sigma_x^2\, \Pi_{p\times d}\boldsymbol{w} = T\sigma_x^2\, \boldsymbol{w}[\mathcal{S}].\end{aligned}$$

Similarly, Term-II(ii) is $\sigma_x^2 \boldsymbol{w}[\mathcal{S}]$. Combining these two, we get Term-II:

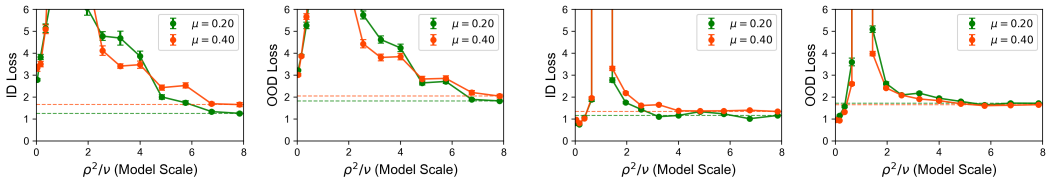
$$\boxed{\mathbb{E}_{X,y}[\mathbf{h}_{\mathcal{S}_p}(X) y_q] = \sigma_x^4 \boldsymbol{w}[\mathcal{S}] \otimes \boldsymbol{w}[\mathcal{S}].}$$

□

H Additional Results and Details of Experimental Settings

Settings. For Fig. 2, we set $d = 80, T = 120, M = 10, n = 400$, and sweep over $p = 2, 6, \ldots, 62$. For Fig. 3, we set $d = p = 80, T = 120, M = 10$, and sweep over $n = 640, 960, 1280, \ldots, 20480$. For the experiments in Figs. 6a and 6b and those in this section, we set $d = 30, T = 40, M = 6, 12, n = 100$, and sweep over $p = 2, 4, \ldots, 28$. In all experiments we set $\sigma_n = 0.1$.

Further, in all experiments, for both ID and OOD loss we use 4000 test sequences each. For OOD evaluation, these come from 4000 fresh task vectors $\boldsymbol{w}$ drawn independently, while for ID the the task vectors $\boldsymbol{w}$ are resampled from the pool of M tasks. Error bars are ± 1 std over 100 runs per point (10 random feature-subset ($\mathcal{S}$) draws $\times$ 10 independent data trials per subset), where each run uses a fresh training set and fresh evaluation sequences. All experiments were run on a Google Colab T4 GPU.



(a) Random Features (b) Top-p Features

Figure 6: **Power Law Covariance**. Experimental (Left) ID and (Right) OOD loss curves where covariance eigenvalues ($\boldsymbol{\Sigma_w}, \boldsymbol{\Sigma_x}$) follow a power-law decay ($\alpha = 1$). (a) Using random feature selection, the overparameterized regime achieves strictly lower risk than the underparameterized minimum (both ID and OOD). (b) Using top-p feature selection, we do not observe in-context benign overfitting; the overparameterized performance is worse than the underparameterized minimum.

Power Law Covariance. Thus far, we have assumed isotropic covariance matrices for $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_w}$, which suffices to establish in-context benign overfitting. For completeness, in Fig. 6a, we replace the isotropic covariance matrices $\boldsymbol{\Sigma_x}$ and $\boldsymbol{\Sigma_w}$ with power-law decay (diagonal entries decaying with exponent $\alpha = 1$; see Appendix H for details). In this setting, we observe a stronger form of in-context benign overfitting: as scale increases, both ID and OOD losses converge to values strictly lower than their respective minima in the underparameterized regime.

In Fig. 6b, instead of random p features, we consider top p features. In this case, consistent with observations in the supervised learning setting [Belkin et al., 2020], we do not observe in-context benign overfitting: As the model scale is increased, both ID and OOD loss values converge to a higher value than in the underparameterized regime.

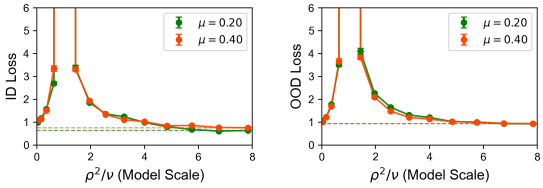


Figure 7: **Power Law Covariance (Only $\boldsymbol{\Sigma_w}$)**. Experimental (Left) ID and (Right) OOD loss curves where task vector covariance eigenvalues follow a power-law decay ($\alpha = 1$). Using random feature selection, the overparameterized regime achieves strictly lower risk than the underparameterized minimum for ID tasks but not for OOD tasks.

Transformer Experiments. In all transformer experiments, we train models to predict the query label given the in-context examples, following our theory setup in Section 2. We set $d = 180, T = 160$, $M = 10, \sigma_n = 0.2$ across all experiments. Models are trained using the Adam optimizer [Kingma and Ba, 2014] using the PyTorch [Paszke et al., 2019] implementation with lr $= 1e-4$, default EMA parameters $\beta_1 = 0.9, \beta_2 = 0.999$, a batch size of 64, and run for 20,000 iterations. For both the linear attention only and full transformer (softmax attention) experiments in Fig. 5, we use 1 head, with $N = 400$ and $N = 800$ respectively. For both ID and OOD loss evaluation, we use 1000 test sequences each. For ID evaluation, task vectors are drawn from $\mathcal{D}_{\text{task}}^{\text{train}}$, whereas for OOD

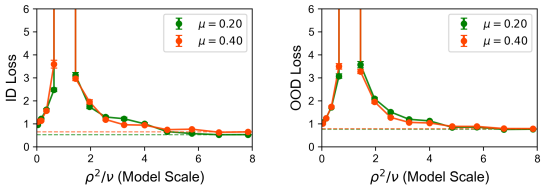


Figure 8: **Power Law Covariance (Only $\boldsymbol{\Sigma_w}$)**. Experimental (Left) ID and (Right) OOD loss curves where task vector covariance eigenvalues follow a power-law decay ($\alpha = 1.5$). Using random feature selection, the overparameterized regime achieves strictly lower risk than the underparameterized minimum for both ID and OOD tasks.

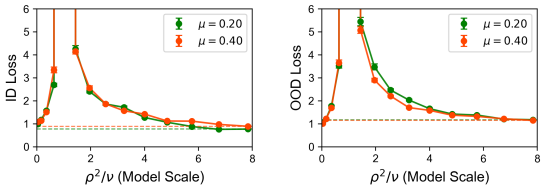


Figure 9: **Power Law Covariance (Only $\boldsymbol{\Sigma_x}$)**. Experimental (Left) ID and (Right) OOD loss curves where feature covariance eigenvalues follow a power-law decay ($\alpha = 1$). Using random feature selection, the overparameterized regime achieves strictly lower risk than the underparameterized minimum for ID tasks but not for OOD tasks.

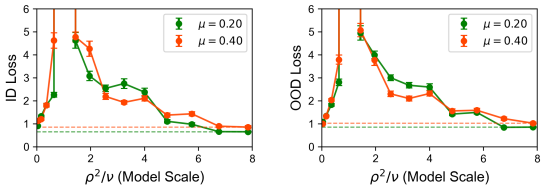


Figure 10: **Power Law Covariance (Only $\boldsymbol{\Sigma_x}$)**. Experimental (Left) ID and (Right) OOD loss curves where feature covariance eigenvalues follow a power-law decay ($\alpha = 2.5$). Using random feature selection, the overparameterized regime achieves strictly lower risk than the underparameterized minimum for ID tasks and for lower μ for OOD tasks.

evaluation, task vectors are sampled independently from $\mathcal{D}_{\text{task}}$. For each model size (embedding dimension), we run 3 independent seeds and report the mean with 95% t-interval shaded bands. We run all experiments on an NVIDIA V100 GPU with 32GB memory.

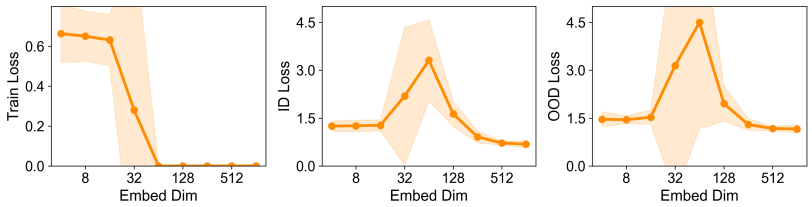


Figure 11: **1 layer linear attention only**. Train loss (left), ID loss (middle), and OOD loss (right) as a function of embedding dimension. The ID and OOD loss curves replicate Fig. 5. The train loss drops to near zero at embedding dimension 64, marking the interpolation threshold, after which the double descent in ID and OOD loss begins.

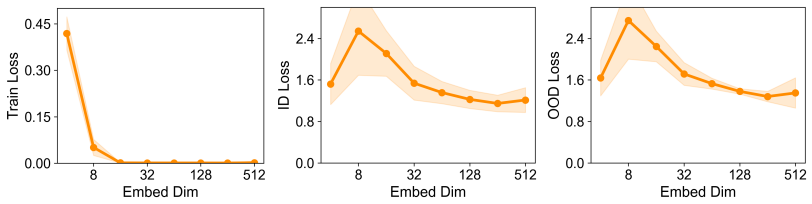


Figure 12: **1 layer full transformer**. Train loss (left), ID loss (middle), and OOD loss (right) as a function of embedding dimension. The ID and OOD loss curves replicate Fig. 5. The train loss drops to near zero at embedding dimension 8, marking the interpolation threshold, after which the double descent in ID and OOD loss begins.

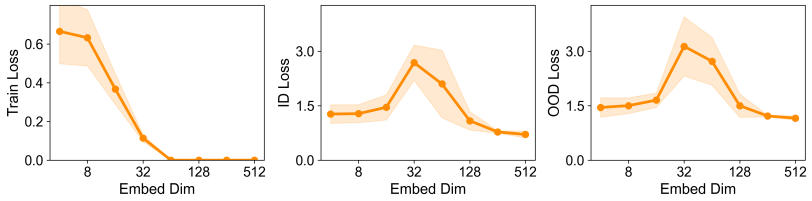


Figure 13: **1 layer linear attention + MLP**. Train loss (left), ID loss (middle), and OOD loss (right) as a function of embedding dimension. We see in-context benign overfitting, similar to Fig. 5, 11.

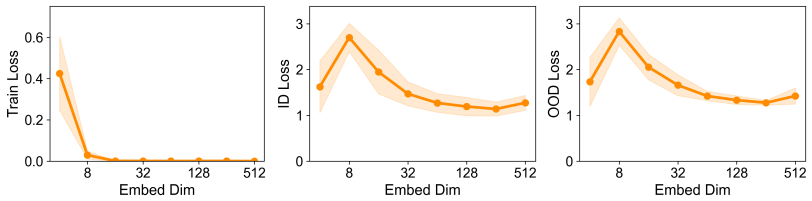


Figure 14: **2 layer full transformer (softmax)**. Train loss (left), ID loss (middle), and OOD loss (right) as a function of embedding dimension. We see in-context benign overfitting, similar to Fig. 5, 12.

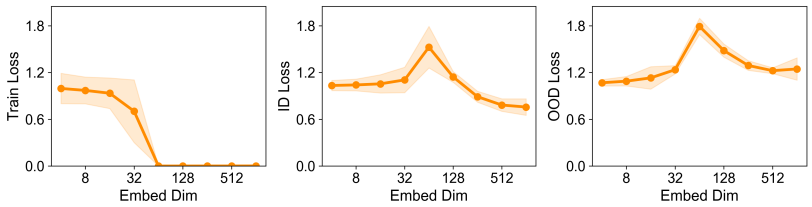


Figure 15: **1 layer linear attention with frozen embedding**. Train loss (left), ID loss (middle), and OOD loss (right) as a function of embedding dimension. Here, the embedding layer weights are frozen and we and only train attention weight matrices. In this case, the model becomes an in-context random features model, analogous to the random features model in supervised learning setting [Hastie et al., 2019]. Similar to results above, ID loss at the overparameterized extreme beats the underparameterized minimum, however the OOD loss is slightly worse.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: See Section 1.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: See Section 6.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: We state the full setup and all assumptions for our theoretical results in Section 2, 3, and discuss proof sketch in Section 5. The full derivations are present in Appendix B to Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: See Section 4 and Appendix H.

   Guidelines:

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

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

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

   Answer: [No]

   Justification: All the experiments conducted in this work are straightforward and small scale, with all details provided in Section 4 and Appendix H.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: See Appendix H.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: See Section 4, Appendix H.

   Guidelines:

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

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See Appendix H.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: N/A.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: N/A.

Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: N/A.

    Guidelines:

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